Réduction de dimension en statistique et application en imagerie hyper-spectrale

Robin Girard

Université Joseph Fourier

THESE

pour obtenir le grade de

DOCTEUR DE L'Université Joseph Fourier

Spécialité : "mathématiques appliquées"

préparée au laboratoire Jean Kuntzmann (LJK)
dans le cadre de **l'Ecole Doctorale "mathématiques, sciences et technologie de l'information, informatique"**

présentée et soutenue publiquement

par

Robin GIRARD

le 26 Juin 2008

Titre :

Réduction de dimension en statistique et application en imagerie hyper-spectrale

---

Directeurs de thèse : Anestis ANTONIADIS et Sophie LAMBERT-LACROIX

JURY

Mme	Valérie PERRIER	, Présidente
M.	Yannick BARAUD	, Rapporteur
M.	Rainer von SACHS	, Rapporteur
M.	Anestis ANTONIADIS	, Directeur de thèse
Mme.	Sophie LAMBERT-LACROIX	, Co-directrice de thèse
M.	Christoph SEGEBARTH	, Examinateur
M.	Yves ROZENHOLC	, Examinateur

Remerciements

Je remercie tout d'abord Madame Valerie Perrier, Professeur à l'Institut National Polytechnique de Grenoble, qui me fait l'honneur de présider mon Jury, ainsi que Christoph Segebarth, Directeur de Recherche à l'INSERM, et Yves Rozehenholc, Maître de conférence à l'Université René Descartes pour leur participation à mon jury de thèse.

Ce travail n'aurait pas pu voir le jours sans Anestis Antoniadis Professeur à l'Université Joseph Fourier, et et Sophie Lambert Lacroix, Maître de conférence à l'Université Joseph Fourier, qui ont dirigé cette thèse. Merci pour la confiance qu'ils m'ont témoigné, les conseils qu'ils m'ont donné et les problèmes scientifiques sur lesquels ils m'ont orientés.

Merci à Clarisse, Paola, Nino et ... . Je suis bien conscient de la chance que cela représente d'avoir une famille à mes côtés qui me donne sans compter (nuit et jour). Ils m'ont fait grandir avec eux et je sais que cela n'a pas de prix. Merci à mes parents, mes frères et soeurs, mes amis du LJK et d'ailleurs, qu'aurais-je fait sans eux ?

Je tiens enfin à remercier Mr Yannick Baraud, Professeur à l'Université de Nice, et Rainer von Sachs de l'Université catholique de Louvain qui ont bien voulu être les rapporteurs de ce travail. En dehors de l'honneur que constitue leur regard sur mon travail, leurs rapports et leurs différentes remarques constructives m'ont permis d'améliorer considérablement certains points de cette thèse.

Résumé

Cette thèse est consacrée à l'analyse statistique de données en grande dimension. Nous nous intéressons à trois problèmes statistiques motivés par des applications médicales : la classification supervisée de courbes, la segmentation supervisée d'images hyperspectrales et la segmentation non-supervisée d'images hyperspectrales. Les procédures développées reposent pour la plupart sur la théorie des tests d'hypothèses (tests multiples, minimax, robustes et fonctionnels) et la théorie de l'apprentissage statistique. Ces théories sont introduites dans une première partie. Nous nous intéressons, dans la deuxième partie, à la classification supervisée de données gaussiennes en grande dimension. Nous proposons une procédure de classification qui repose sur une méthode de réduction de dimension et justifions cette procédure sur le plan pratique et théorique. Dans la troisième et dernière partie, nous étudions le problème de segmentation d'images hyper-spectrales. D'une part, nous proposons un algorithme de segmentation supervisée reposant à la fois sur une analyse multi-échelle, une estimation par maximum de vraisemblance pénalisée, et une procédure de réduction de dimension. Nous justifions cet algorithme par des résultats théoriques et des applications pratiques. D'autre part, nous proposons un algorithme de segmentation non supervisée impliquant une décomposition en ondelette des spectres observées en chaque pixel, un lissage spatial par croissance adaptative de régions et une extraction des frontières par une méthode de vote majoritaire.

Mots-clés : segmentation, traitement d'images, images hyper-spectrales, imagerie médicale, détection de contours, transformées en ondelettes, réduction de dimension, Analyse discriminante linéaire et quadratique, données fonctionnelles, maximum de vraisemblance pénalisée, mixlet, Lissage adaptatif, perturbation de règle de décision.

Abstract

This thesis deals with high dimensional statistical analysis. We focus on three different problems motivated by medical applications : curve classification, pixel classification and clustering in hyperspectral images. Our approaches are deeply linked with statistical testing procedures (multiple testing, minimax testing, robust testing, and functional testing) and learning theory. Both are introduced in the first part of this thesis. The second part focuses on classification of High dimensional Gaussian data. Our approach is based on a dimensionality reduction, and we show practical and theorical results. In the third and last part of this thesis we focus on hyperspectral image segmentation. We first propose a pixel classification algorithm based on multi-scale analysis, penalised maximum likelihood and feature selection. We give theorical results and simulations for this algorithm. We then propose a pixel clustering algorithm. It involves wavelet decomposition of observations in each pixel, smoothing with a growing region algorithm and frontier extraction based on a voting scheme.

Table des matières

Introduction

Les technologies d'acquisition de données sont de plus en plus performantes et les stockages effectués sont de plus en plus systématiques. Ceci a amené la communauté statistique à étudier des problèmes dans lesquels, le nombre d'observations n est nettement plus petit que le nombre p d'attributs de ces observations. Les problèmes statistiques d'étude de données de ce type sont du domaine de la statistique en grande dimension. Cette thèse contribue à l'étude de ce type de problèmes. Elle a été financée par la région Rhônes-Alpes dans le cadre d'un projet région sur la thématique cancer et c'est donc, d'une certaine façon, l'évolution des données en cancérologie qui a motivé notre travail. Comme nous le verrons par la suite, les motivations pratiques vont au delà des seules applications médicales.

Dans le cadre du projet région cité ci-dessus, nous avons travaillé avec une équipe de chercheurs de l'unité mixte INSERM/UJF 594 nouvellement unité INSERM U836. Cette équipe de médecin-physiciens cherche à mettre en oeuvre un logiciel permettant d'assister le bilan préopératoire dans le diagnostic des tumeurs cérébrales. Les histopathologistes distinguent en général trois familles de tumeurs cérébrales : les tumeurs primitives gliales (ou gliomes), les tumeurs primitives non gliales, et les métastases. Ces distinctions sont basées sur le type de la tumeur, c'est-à-dire sur la nature des cellules devenues cancéreuses. Par exemple les métastases ne sont pas, comme les deux autres, des tumeurs primitives issues de cellules cérébrales. Elles sont issues de tumeurs secondaires provenant de cellules tumorales d'un autre organe qui ont essaimé vers le cerveau. Dans chaque type de tumeur on distingue ensuite un grade (degré de malignité) qui va de I à V. Un diagnostic consiste en l'identification d'une tumeur pour la prescription d'un traitement adapté. Rappelons qu'il s'effectue en deux temps. Le premier consiste en un bilan préopératoire destiné à obtenir un diagnostic préliminaire. Celui-ci va guider le choix thérapeutique et préparer une possible intervention chirurgicale. Ce bilan est obtenu principalement grâce aux données cliniques (âge, symptômes, antécédents, présence éventuelle d'une autre tumeur...) et aux données d'imagerie (scanner, IRM). La deuxième étape du diagnostic consiste en une ou deux biopsies. Une biopsie comporte des risques et les renseignements préopératoires peuvent être des éléments déterminants vis à vis de la qualité du diagnostic final et de l'utilité de la biopsie. Il est donc intéressant d'améliorer les techniques non invasives existantes pour fournir un diagnostic préopératoire plus précis.

Nous cherchons dans ce mémoire à mettre en oeuvre des méthodes statistiques permettant d'améliorer la première phase du bilan préopératoire. Notre étude se focalise sur un type nouveau de données d'imagerie : les images spectroscopiques.

L'imagerie spectroscopique (voir [20] pour une introduction à l'imagerie par résonance magnétique en général) est la forme d'imagerie médicale qui consiste en l'acquisition simultanée

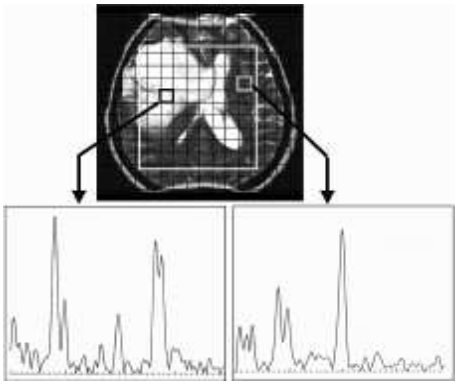


FIG. 1 – Le spectre observé sur chaque pixel caractérise les fréquences propres de vibration des différents tissus qui composent le pixel.

de l'ensemble des spectres localisés dans un plan de coupe du cerveau. Chacun de ces spectres est la réponse en fréquence des différentes substances chimiques qui composent le tissu compris dans un volume élémentaire (voxel) du plan de coupe considéré (voir Figure 1). Il permet aux médecins spécialistes de la spectroscopie de présumer un type histopathologique de tumeur. En chaque pixel de l'image (associé à un volume élémentaire) on observe donc une courbe échantillonnée en un grand nombre de points. On parle aussi d'image hyper-spectrale. Ce type d'image est également très courant hors du champ médical, notamment dans le domaine de l'imagerie satellite.

Les études existantes en matière de traitement de signaux issus d'examen IRM n'utilisent pas souvent la caractérisation spectroscopique des tissus. Le peu d'études de ce type ayant été effectuées ne s'attachent qu'au spectre correspondant à un voxel dont on sait à l'avance qu'il est issu d'une tumeur dont on veut connaître les propriétés. Cette démarche est bien évidement limitée. Il est parfois difficile de connaître exactement la localisation d'une tumeur mais surtout il est très rare de pouvoir isoler toute l'information d'une tumeur dans un voxel. D'un côté, en un voxel donné, une tumeur est un mélange de tissus plus ou moins malades, et de l'autre, en deux voxels différents d'une même tumeur, plusieurs tissus différents peuvent cohabiter. Aussi, nous allons chercher à caractériser une tumeur donnée en intégrant l'information comprise dans toute l'image.

Depuis plusieurs années un certain nombre de travaux très intéressants ont cherché à identifier des tumeurs au sein d'une image hyper-spectrale (voir par exemple Szabo De Edeleneyi [73] ou Hagberg [40]). Les méthodes qui y sont développées se décomposent en deux étapes. La première consiste en l'apprentissage du lien qui existe entre un spectre et la tumeur à laquelle il est associé. Ce premier temps repose donc sur la constitution d'un échantillon d'apprentissage, c'est-à-dire la donnée d'une série de spectres et des tumeurs associées. Le deuxième temps est la segmentation d'une image sur laquelle rien n'est connu, c'est-à-dire l'identification en chaque pixel du type de tissus et éventuellement de la tumeur associée. Ces méthodes comportent trois limites auxquelles nous voulons apporter des solutions.

Tout d'abord, les données enregistrées ne sont pas traitées comme des données fonctionnelles (i.e. des courbes). Autrement dit ces travaux se limitent à un choix a priori de certains attributs caractéristiques des spectres observés. Le nombre d'attributs et le choix de ces attributs n'est

donc pas guidé par les données elles-mêmes. Les résumés vectoriels finalement choisis sont utilisés pour discriminer ou prédire le type histopathologique des tissus avec des techniques classiques de reconnaissance de forme ou d'analyse statistique multidimensionnelle. Les indicateurs (attributs) le plus souvent utilisés sont des bandes spectrales, dont on sait qu'elles apportent une information pertinente. Dans sa thèse, Szabo De Edeleneyi [73] utilise la proportion d'aire comprise sous sept bandes spectrales pertinentes pour construire sept attributs. Le nombre sept semble être choisi comme permettant de recouvrir les bandes spectrales pertinentes et non comme une dimension idéale pour la discrimination. Il nous a semblé naturel et souhaitable de chercher à utiliser des méthodes de classification et de segmentation exploitant toute l'information disponible. Ainsi, nous voulons être en mesure de sélectionner de manière automatique un certain nombre d'attributs. Nous voulons être capable de juger dans quelle mesure cela est nécessaire pour améliorer les performances de classification.

Ensuite, la plupart des techniques utilisées sont basée sur l'utilisation d'un échantillon d'apprentissage. Cet échantillon d'apprentissage est composé de spectres observés sur différents patients et différentes tumeurs que l'on a au préalable identifié. La construction de ces échantillons d'apprentissage nécessitent un prétraitement comprenant l'extraction d'une zone de l'image où la tumeur est présente. Cette extraction n'est pas faite sur l'image spectroscopique, mais sur une image plus facile à visualiser (pondérée $T1$ ou $T2$). Elle est bien évidemment coûteuse en temps et en main d'oeuvre qualifiée, et il est naturel de vouloir automatiser ce prétraitement.

Enfin, lors de la segmentation, la cohérence a priori de deux voxels proches n'est pas prise en compte. En effet, si l'échantillon d'apprentissage est obtenu à partir de plusieurs pixels pour une image donnée, un algorithme de segmentation appliqué à une nouvelle image interprète chaque pixel indépendamment des autres. Le problème traité par un tel algorithme est donc exactement un problème de classification de pixels. Nous voulons donner une approche du problème de segmentation qui tienne compte, par une modélisation adéquate, de ce qui fait la différence entre un problème de segmentation et un problème de classification.

Pour répondre à certaines attentes des médecins, nous avons voulu mettre en oeuvre trois méthodes utilisant toute l'information contenue dans les spectres et les images. Nous avons choisit de traiter d'abord le problème associé à la première limitation et posé par la classification de données fonctionnelles. Les deux autres problèmes que nous nous posons sont des problèmes de segmentation. Nous insistons sur le fait que le problème de segmentation peut être traité par une méthode de classification, mais nous voulons en plus tenir compte de la similarité a priori entre deux tissus associés à deux pixels voisins. Dans ce cadre, et en utilisant le caractère fonctionnel des données, nous voulons traiter deux problèmes : le problème de segmentation non supervisée (pour donner une solution à la deuxième limitation) et le problème de segmentation supervisée. Avant de donner le plan de ce mémoire, nous allons définir dans le détail ces problèmes. Nous allons donner le formalisme mathématique associé à ces problèmes, et insister sur les difficultés posées par ceux-ci.

Le formalisme mathématique associé au problème consistant à prédire un type histopathologique à partir d'un spectre est celui de la classification. La classification consiste à prédire la nature y, appelée aussi classe ou label, d'une observation x. Dans le cas le plus simple, celui de la classification binaire, le label prend ses valeurs dans $\{0, 1\}$, et dans le cas de la classification à K classes, y prend ses valeurs dans $\{1, \ldots, K\}$. L'observation x est très souvent constituée d'un certain nombre d'attributs numériques formant un vecteur de $\mathcal{X} = \mathbb{R}^p$, mais elle peut aussi

être une courbe ou une image. La classification en dimension finie est le cas où $\mathcal{X} = \mathbb{R}^p$ et la classification de courbes est le cas où $\mathcal{X}$ est un espace fonctionnel éventuellement de dimension infinie.

En classification à K classes, on construit une application g de $\mathcal{X}$ dans $\{1, \dots, K\}$ qui à une observation $x \in \mathcal{X}$ associe la prédiction faite. Cette application est une fonction de décision que l'on appelle classificateur. Ce classificateur commet une erreur dans la prévision associée à x si $g(x) \neq y$.

Pour formaliser le problème d'apprentissage tel que nous l'envisageons, il faut introduire un formalisme probabiliste. Ainsi, nous supposons que (X, Y) est une variable aléatoire à valeur dans $\mathcal{X} \times \{1, \dots, K\}$ modélisant les observations et les classes associées. Pour $k \in \{1, \dots, K\}$, nous notons P_k la loi de $(X|Y = k)$. Cette loi modélise la distribution des observations issues de la classe k. On souhaite naturellement construire un classificateur performant, c'est-à-dire qui se trompe avec une probabilité la plus faible possible. Il existe dans ce problème K erreurs de natures différentes consistant à ne pas affecter une observation à la classe $k \in \{1, \dots, K\}$ alors que son label vaut effectivement k. Pour k fixé, l'erreur est mesurée par $P_k(g(X) \neq k)$. Il y a alors plusieurs approches selon l'importance que l'on donne aux divers types d'erreurs. L'approche bayesienne est la plus courante. Elle consiste à considérer l'erreur de classification $P(g(X) \neq Y)$. Le classificateur qui minimise cette probabilité d'erreur est la règle de Bayes. Cette erreur peut être réécrite grâce à la formule de Bayes :

$$\mathcal{C}(g) = P(g(X) \neq Y) = \sum_{k=1}^{K} P(Y = k) P_k(g(X) \neq k). \tag{1}$$

Notons que cette mesure d'erreur donne une grande importante à $P_k(g(X) \neq k)$ si la classe k a une forte probabilité d'apparition. En d'autres termes, dans le cadre bayesien, l'erreur $P_k(g(X) \neq k)$ associée à une affectation erronée d'une observation de la classe k, a d'autant plus d'importance que le label k apparaît avec une grande probabilité. Dans notre cadre, il n'est pas raisonnable de penser que la fréquence d'apparition d'une tumeur ayant un label $k \in \{1, \dots, K\}$ nous renseigne sur l'importance que l'on doit donner à l'erreur faite lorsque $g(X) \neq k$ sachant que Y vaut k. En effet, il n'est pas naturel d'attribuer à une tumeur apparaissant peu souvent une faible importance tant il est vrai que dans le domaine médical la rareté est souvent synonyme de pathologie sérieuse. Nous supposons donc que Y est distribué de manière uniforme sur $\{1 \dots, K\}$. Dans ce cas, l'erreur de classification associée à un classificateur g (définie par (1)) est

$$\mathcal{C}(g) = \frac{1}{K} \sum_{k=1}^{K} P_k(g(X) \neq k). \tag{2}$$

Dans tous nos problèmes de classification nous supposerons que Y suit une loi uniforme.

Dans le cas où les lois P_k sont connues, la règle de Bayes g^* est entièrement déterminée. Dans la réalité, ces lois ne sont pas connues mais on dispose d'un échantillon d'apprentissage, composé pour toute classe $k \in \{1, \dots, K\}$, de n_k observations $X_1^k, \dots, X_{n_k}^k$ de variables aléatoires indépendantes et identiquement distribuées selon la loi P_k. Le problème consistant à construire un classificateur $\hat{g}$ à partir de ces données est presque centenaire. La première approche à ce problème remonte à 1936 et fut proposée par Fisher [34]. L'approche de Fisher est liée à la procédure LDA (Linear Disciminant Analysis). Dans la procédure LDA, les distributions $(P_k)_{k=1,\dots,K}$ sont supposées gaussiennes de moyennes différentes et de covariances égales. Dans ce cas, la règle de

Bayes (celle qui minimise (2)) s'exprime simplement en fonction des moyennes et de la covariance commune des lois $(P_k)_{k=1,\dots,K}$. La procédure LDA consiste alors à estimer ces paramètres et à imiter la règle de Bayes avec ces estimations. L'analyse quadratique discriminante (QDA) est basée sur le même principe mais dans le cas où les matrices de covariance des différents groupes sont différentes. Notons qu'il est aussi courant de parler de procédure LDA et QDA lorsque Y n'est pas uniformément distribué. Dans notre cas (Y suit une loi uniforme), la procédure LDA obtenue avec les estimateurs empiriques des moyennes et de la covariance commune des différents groupes correspond exactement à la procédure de Fisher. Notons enfin qu'une procédure, comme LDA et QDA, consistant à imiter la règle optimale, en introduisant une estimation des paramètres qui la définissent, est une procédure de type plug-in.

Un classificateur obtenu par la procédure LDA sera noté g^{LDA} et un classificateur obtenu par la procédure QDA sera noté g^{QDA}. Dans le cas de la classification binaire, une règle de classification g sépare l'espace des observations en deux parties : l'une dans laquelle g vaut 0 et l'autre dans laquelle g vaut 1. Dans le cas de la procédure LDA, les deux parties associées au classificateur g^{LDA} sont délimitées par un hyperplan affine. Dans le cas de la procédure QDA, les deux parties associées aux classificateurs g^* et g^{QDA} sont délimitées par une forme quadratique. Le classificateur g^{LDA} sera un nom générique pour désigner un classificateur défini par un hyperplan affine. De même g^{QDA} sera un nom générique pour parler d'un classificateur défini par une forme quadratique. Nous parlerons de règle de Fisher lorsque g^{LDA} sera construit à partir de la moyenne empirique et de la covariance empirique.

Il est naturel de vouloir mesurer la différence entre la règle g^* et, selon la modélisation, la règle g^{LDA} ou g^{QDA}. D'une manière générale, lorsque g^* est le classificateur optimal pour l'erreur $\mathcal{C}(.)$ (définie par (1)), on définit l'excès de risque associé à un classificateur g par :

$$\mathcal{C}(g) - \mathcal{C}(g^*). \tag{3}$$

Cette dernière quantité vaut

$$P(g(X) \neq Y \text{ et } g^*(X) = Y) - P(g^*(X) \neq Y \text{ et } g(X) = Y). \tag{4}$$

Dans le cas des procédures LDA et QDA, nous avons choisi d'étudier ce que nous avons appelé erreur d'apprentissage. Cette erreur est définie par la probabilité d'effectuer une erreur de classification avec la règle g alors que la règle optimale g^* n'en fait pas :

$$\mathcal{R}(g) = P(g(X) \neq Y \text{ et } g^*(X) = Y). \tag{5}$$

Notons que si g est construit à partir de l'échantillon d'apprentissage, ce terme d'erreur est alors une variable aléatoire mesurable par rapport à l'échantillon d'apprentissage. Notons que les quantités définies par (3) et (5) sont différentes et que l'erreur d'apprentissage (5) est un majorant de (3) :

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq \mathcal{R}(g).$$

Plaçons nous pour simplifier dans le cas où $K = 2$ et $Y \rightsquigarrow \mathcal{U}(\{1, 2\})$. Rappelons que $d_1(P_1, P_2) = \int |dP_1 - dP_2| = (1 - 2\mathcal{C}(g^*))$ (voir Partie I Chapitre 1). Dans le cas gaussien, nous avons obtenu le résultat suivant (Chapitre 5 Partie I). Si $d_1(P_1, P_2) \geq C > 0$, alors il existe une constante $C' > 0$ (dépendant de $C > 0$ uniquement) telle que $\mathcal{C}(g) - \mathcal{C}(g^*) \geq C'\mathcal{R}(g)^8$. Ainsi, dans ce cas, on ne peut pas construire une suite g_n de classificateurs telle que $\mathcal{C}(g_n) - \mathcal{C}(g^*)$ tend vers zéro alors que l'erreur d'apprentissage $\mathcal{R}(g_n)$ ne tend pas vers zéro. Par ailleurs,

$d_1(P_1, P_2) = \int |dP_1 - dP_2|$ est proche de zéro exactement lorsque l'on ne peut pas prédire Y avec X, c'est-à-dire lorsque X et Y sont indépendants. Dans, ce cas, puisque

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq d_1(P_1, P_2)$$

Ainsi, pour l'étude des procédures LDA et QDA, les deux critères ne diffèrent singulièrement que lorsque X et Y sont quasiment indépendants. Dans le cas de la procédure LDA par exemple, supposons que $\mathcal{X} = \mathbb{R}$, et que les observations des deux groupes sont issues de lois ayant deux moyennes μ_1 et μ_2 telles que $\mu_1 = -\mu_2$. Alors, la règle optimale correspond à regarder le signe de la nouvelle observation X. Si $\mu_1 \approx \mu_2 \approx 0$, une règle qui déciderais systématiquement de l'appartenance de X au groupe 2 serait à peut près égale (au sens de l'excès de risque) à la règle optimale, mais nettement différente selon l'erreur d'apprentissage.

Les deux critères mentionnés ont leurs qualités et leur défauts. Cette « indulgence » de l'excès de risque lorsque toutes les règles se valent ne permet pas de discréditer une méthode de classification de manière uniforme. L'erreur d'apprentissage n'a pas ce défaut, mais crée un problème délicat, imiter la meilleure règle dans le cas où Y et X sont quasiment indépendantes, alors que le problème est simple. En définitive, nous avons choisi d'utiliser, dans le cadre du problème de classification de données gaussiennes, l'erreur d'apprentissage. Elle nous permet de discréditer certaines méthodes d'apprentissage (par exemple la règle de Fisher en grande dimension) et autorise des interprétations ensemblistes intéressantes. En effet, pour obtenir $\mathcal{R}(g)$, il suffit de calculer la mesure dans l'espace des observations de l'ensemble des points qui sont bien classés par g^* et mal par g.

L'étude des performances pratiques et théoriques d'algorithmes de classification ou plus généralement d'apprentissage a connu ces quinze dernières années un grand succès. Cet essor doit beaucoup à la communauté du « Learning » à l'interface entre Mathématiques Informatique et Sciences Cognitives. A la fin des années 70, Vapnik propose une stratégie générale pour la construction d'un classificateur à partir d'un échantillon d'apprentissage. Cette stratégie, basée sur la minimisation du taux d'erreur empirique, est appelée stratégie ERM (pour Empirical Risk Minimization). Le taux d'erreur empirique est une version empirique du risque défini par (1) et est obtenu grâce à l'échantillon d'apprentissage. L'apprentissage et la classification par plug-in diffèrent de l'approche de Vapnik et ne connaissent pas un succès de même envergure. Dans un cadre assez général, les propriétés des classificateurs par plug-in ont été remises en cause par Yang [77]. Le cadre dans lequel l'étude de Yang se place est vraisemblablement trop général. Il jette un discrédit théorique sur des méthodes qui, dans la pratique, ont souvent fait preuve de leur efficacité. La restriction induite par l'hypothèse de marge introduite par Tsybakov (voir par exemple Audibert et Tsybakov [6] et les références qui y sont faites) permet de montrer que, sous certaines conditions naturelles, les règles plug-in ont de très bonnes performances théoriques. Cependant le cadre théorique qui est discuté par Audibert et Tsybakov [6] n'est pas particulièrement adapté à l'étude des procédures LDA et QDA, et ne donne pas de réponse aux problèmes posés par la grande dimension telle que nous l'envisageons.

Dans le cas où les observations appartiennent à un espace de petite dimension p, il n'existe pas de résultats théoriques convaincants permettant de comparer de manière précise les performances de la procédure optimale avec celles des procédures de type g^{LDA} et g^{QDA}. Qui plus est, le problème que nous nous posons est d'autant plus difficile qu'il s'agit de fabriquer, à partir de l'échantillon d'apprentissage, un classificateur qui ait des bonnes performances pratiques et théoriques dans le cadre de la grande dimension. Autrement dit, il faut être capable de construire et

défendre une procédure lorsque la dimension p de l'espace des observations est bien plus grande que $n = \sum_{k=1}^{K} n_k$ le nombre total d'observations de l'échantillon d'apprentissage. Dans ce cas, la règle de Fisher est connue pour ses mauvaises performances pratiques. Bickel et Levina [12] donnent une explication théorique à cela.

Lorsque l'on travaille sur des données de grande dimension ($n \ll p$), il n'est pas toujours pertinent de prédire le label à partir de l'ensemble des attributs disponibles comme le fait la règle de Fisher. En effet, dans le cas où seulement un petit nombre d'attributs contribue à la séparation des données, la présence d'une trop grande quantité d'information superflue peut amener à un sur-ajustement, c'est-à-dire à la construction d'un classificateur qui ne soit efficace que sur l'échantillon d'apprentissage et pas sur les nouvelles données. Il est donc intéressant de construire un classificateur qui soit capable, dans ce cas, de détecter au préalable la quantité d'information pertinente. Une telle procédure de détection correspond à une réduction de dimension. Récemment, Fan et Fan [32] ont proposé une procédure de réduction de dimension basée sur une procédure de test. D'un point de vu théorique, ils montrent que leur procédure de réduction de dimension choisit, lorsque le nombre d'attributs est assez grand, les attributs qui contribuent effectivement à la séparation des données. Ils ne montrent pas dans quelle mesure cette procédure permet d'obtenir un bon classificateur. Il est donc pertinent de chercher à contrôler l'erreur d'apprentissage (et donc l'excès de risque défini par (3)) associée aux classificateurs g^{LDA} et g^{QDA}. Il est intéressant, pour permettre un choix adapté de la procédure d'estimation des paramètres définissant g^{LDA} et g^{QDA}, que ce contrôle soit indépendant de la procédure d'apprentissage utilisée. Ceci n'a encore jamais été envisagé, Bickel et Levina [12] donnent, pour une procédure bien particulière, le lien entre l'espérance de l'erreur de classification et le pire des risques de Bayes sur une classe de problèmes. Notons que ceux-ci n'utilisent pas de procédure de réduction de dimension.

Les procédures de type QDA et LDA sont des procédures de classification parmi les plus vieilles qui existent. Elles sont encore, ainsi que quelques variantes, parmi les plus utilisées dans la pratique. La principale critique faite à ces procédures est l'hypothèse gaussienne qu'elles nécessitent. Ainsi, le manque de généralité de ces procédures semble être à l'origine du désintérêt pour elles, des théoriciens. Ceci est certainement une des raisons à l'absence de résultats théoriques puissants les concernants. Pourtant, elles offrent un cadre simple et assez robuste qui donne libre court à certaines interprétations géométriques éclairantes, et dans d'autre problèmes, le cadre gaussien a très souvent donné naissance à des procédures et idées théoriques intéressantes.

En définitif, les méthodes de classification de données gaussiennes en grande dimension restent très utilisées dans la pratique, mais ont été très peu étudiées sur le plan théorique. Nous cherchons dans ce mémoire à donner un éclairage nouveau et plus avancé à ce problème.

Le formalisme et la problèmatique de la segmentation d'images hyper-spectrales héritent de ceux de la classification en grande dimension. La différence entre les deux problème sera explicité et exploitée plus loin. Rappelons qu'une image peut être modélisée par un ensemble structuré $\mathcal{T}_N$ de N pixels auxquels sont associées des observations $(x_i)_{i\in\mathcal{T}_N}$ à valeur dans un espace $\mathcal{X}$. On parle d'image hyper-spectrale lorsque $\mathcal{X}$ est un espace de grande dimension ou de dimension infinie. On parle d'image multidimensionnelle lorsque $\mathcal{X}$ est de dimension plus grande que 1 mais que cette dimension reste faible. Dans la suite, une image désignera de manière indifférente un de ces types d'images. La segmentation d'image consiste à prédire les labels $(y_i)_{i\in\mathcal{T}_N}$, associées aux observations $(x_i)_{i\in\mathcal{T}_N}$. Dans le cas le plus simple, celui de la segmentation binaire, un label prend ses valeurs dans $\{0, 1\}$, et dans le cas de la segmentation à K classes un label y_i prend ses

valeurs dans $\{1, \ldots, K\}$.

En segmentation à K classes, on construit une application h de $\mathcal{X} \times \mathcal{T}_N$ dans $\{1, \ldots, K\} \times \mathcal{T}_N$ qui, à une observation $x \in \mathcal{X}$ en un pixel $i \in \mathcal{T}_N$, associe une prédiction. Cette application est une fonction de décision que l'on appelle fonction de segmentation. Cette fonction de segmentation se trompe sur l'observation x au pixel i si $h(x, i) \neq y_i$.

Pour formaliser le problème de segmentation, de la même manière que pour le problème de classification, il faut introduire un cadre probabiliste. Ainsi, nous supposons que $(X_i, Y_i)_{i \in \mathcal{T}_N}$ est une famille de variables aléatoires indépendantes à valeurs dans $\mathcal{X} \times \{1, \ldots, K\}$ modélisant les observations sur l'image et les classes associées. En un pixel $i \in \mathcal{T}_N$, la loi de X_i est notée P^i et nous notons P_k la loi de $(X_i | Y_i = k)$. Cette loi modélise la distribution des observations issues de la classe k, et, comme la notation l'indique, nous supposons que cette loi ne dépend pas de la position spatiale $i \in \mathcal{T}_N$. On souhaite naturellement construire une fonction de segmentation performante, c'est-à-dire telle que l'espérance du nombre de pixels mal classés :

$$\mathcal{E}(h) = \mathbb{E}\left[\sum_{i \in \mathcal{T}_N} 1_{h(X_i, i) \neq Y_i}\right], \tag{6}$$

soit la plus petite possible. La règle optimale, celle qui minimise cette espérance, est donnée par la règle de Bayes h^* qui s'exprime en fonction de $(P_i)_{i \in \mathcal{T}_N}$.

Sans hypothèse supplémentaire, le problème de segmentation est exactement équivalent à N problèmes identiques de classification. L'intérêt d'un problème de segmentation est donc de rajouter une hypothèse modélisant la cohérence spatiale entre les différents pixels de l'image, et de mettre en oeuvre une procédure qui permette de tirer parti de cette information. Dans ce but, nous supposons que l'ensemble $\mathcal{T}_N$ des pixels de notre image hyper-spectrale peut être découpé en M régions homogènes $\{D_1, \ldots, D_M\}$. L'homogénéité de ces régions peut alors être modélisée en supposant que, dans une région donnée D_m, l'application qui à $i \in \mathcal{T}_N$ associe $\pi_{ik} = P(Y_i = k)$ est constante et que le nombre de pixels utilisés pour recouvrir la frontière entre les différentes zones est suffisamment petit. Cette dernière hypothèse est une hypothèse de régularité de frontière.

Rappelons que la connaissance des lois $(P_k)_k$ et des poids π_{ik} permet de construire la règle de Bayes h^*. Nous allons fabriquer une règle de segmentation de type plug-in, c'est-à-dire qui est construite de la même manière que la règle de Bayes, mais avec une estimation des poids π_{ik} et des densités P_k. Pour réaliser notre estimation des poids et des densités, nous avons envisagé deux types de modélisations. Dans la première, nous n'imposons pas de restriction supplémentaire sur les poids $(\pi_{ik})_{i \in \mathcal{T}_N, k \in \{1, \ldots, K\}}$ et supposons seulement que les régions $\{D_1, \ldots, D_M\}$ sont la discrétisation de régions séparées par des frontières régulières. Ainsi, dans cette approche,

$$P^i = \sum_{k=1}^{K} \pi_{ik} P_k \;\text{ et } \forall m \in \{1, \ldots, M\},\;\; (i, j) \in D_m^2, \quad P^i = P^j. \tag{7}$$

Dans la deuxième modélisation, on suppose que dans une zone D_m donnée, une classe seulement peut apparaître. En d'autres termes :

$$\exists k \in \{1, \ldots, K\} \;\; : \;\; \forall i \in D_m, \;\; P(Y_i = k) = 1.$$

Dans les deux cas nous supposons que la loi P_k est donnée par une loi gaussienne sur l'espace $\mathcal{X}$ de grande dimension. Dans le premier cas, les paramètres de ces lois sont estimés grâce à un

échantillon d'apprentissage, et l'on parle alors de segmentation supervisée. Dans le deuxième cas, la covariance de ces lois est connue à une constante d'échelle près et les moyennes de ces lois sont inconnues. Dans cette deuxième approche, on ne possède pas d'échantillon d'apprentissage. On parle alors de segmentation non supervisée. Dans tous les cas, on construit une fonction de segmentation $\hat{h}$ que l'on va chercher à comparer avec la règle optimale h^* (celle minimisant (6)). Pour cela, on peut définir, de la même manière que dans le cas de la classification, l'excès de risque comme étant

$$\mathcal{S}(\hat{h}) = \mathbb{E}\left[\sum_{i\in\mathcal{T}_N} 1_{\hat{h}(X_i,i)\neq Y_i}\right] - \mathbb{E}\left[\sum_{i\in\mathcal{T}_N} 1_{h^*(X_i,i)\neq Y_i}\right], \tag{8}$$

la première espérance étant prise par rapport aux observations de l'image et éventuellement de l'échantillon d'apprentissage. Notons que contrairement à ce que nous faisons dans le problème de classification supervisée, nous tenons compte ici de la fréquence d'apparition des différentes classes. La raison est simple : cette fréquence modélise ici la présence dans l'image d'un type de spectre. L'homogénéité de l'image est exactement ce qui peut être utilisé dans un problème de segmentation et qui n'existe pas dans un problème de classification. La régularité imposée sur les frontières séparant les différentes zones de l'image, dans lesquelles les fréquences d'apparition des différentes classes sont supposées constantes, est exactement ce dont nous cherchons à tirer parti.

La recherche des poids du mélange, dans les cas supervisés et non supervisés, nous met face aux mêmes difficultés que dans le problème de classification en grande dimension. Il est donc préférable de chercher à associer une procédure de réduction de dimension à l'apprentissage des distributions modélisant les différents types de spectres. Cette procédure de réduction de dimension doit être intégrée dans un algorithme tenant compte de la régularité des frontières entre les différentes zones de l'image. Par ailleurs, si dans le cas supervisé, la démarche est en beaucoup de points semblable à la classification supervisée, l'intégration d'une procédure de réduction de dimension pose quelques problèmes supplémentaires.

Dans le cadre de la classification ou de la segmentation non supervisée, l'intérêt d'une procédure de réduction de dimension ne peut pas être expliqué de la même manière que dans le cas de la classification ou de la segmentation supervisée. Il s'agit donc d'identifier l'origine du problème posé par la grande dimension et d'y apporter des solutions.

Dans nos problèmes de classification et de segmentation, nous tenons compte du caractère fonctionnel des observations. Ainsi, nous intégrons des connaissances a priori sur la régularité des courbes observées.

Plan et description de ce mémoire

Ce document comporte trois parties dont nous allons décrire le contenu. Dans la première partie, nous rappelons des éléments de la théorie de la décision et des tests d'hypothèses. Nous abordons ainsi les procédures concernant les hypothèses simples, les problématiques minimax et bayesiennes, les tests d'hypothèses multiples et la théorie de la classification par plug-in. Dans la deuxième partie, nous abordons le problème de classification de courbes. Nous travaillons dans un formalisme gaussien et nous utilisons une règle plug-in (type LDA ou QDA) basée sur une méthode de réduction de dimension. Cette méthode repose sur une étude théorique de l'influence

de l'erreur d'estimation des paramètres sur l'erreur d'apprentissage. La dernière partie présente une méthode de segmentation supervisée d'images hyper-spectrales et une méthode non supervisée ainsi que les résultats théoriques et pratiques liés à celles-ci.

La première partie de ce mémoire est consacrée aux tests et aux problèmes d'apprentissage. Cette partie ne contient pas de résultats singulièrement nouveaux, mais quelques formulations originales. Nous y donnons entre autres des introductions utiles à la compréhension de certaines notions théoriques et pratiques de ce mémoire.

Sur le plan théorique, le formalisme des tests d'hypothèses constitue un outil puissant et particulièrement adapté aux questions que nous nous posons dans cette thèse. Rappelons que si $\mathcal{X}$ est un espace d'observations, sur lequel deux ensembles de mesures de probabilités disjoints $\mathcal{P}_0$ et $\mathcal{P}_1$ sont définis, un test de l'hypothèse H_0 contre H_1 est une fonction de $\mathcal{X}$ dans $\{0,1\}$ permettant de décider si une observation $X \in \mathcal{X}$ est issue d'une loi de $\mathcal{P}_0$ ou d'une loi de $\mathcal{P}_1$. Ainsi, un problème de segmentation ou de classification binaire peut être formulé et modélisé par des tests. Il suffit remplacer le mot « hypothèses » par « classes » . Remarquons que les cas de classifications et de segmentations à K classes se traitent aussi dans le formalisme des tests.

Dans un problème de classification supervisée, si l'échantillon d'apprentissage est infini, on a la possibilité de connaître parfaitement les lois P_0 et P_1. Les hypothèses correspondantes sont alors simples. Le formalisme des tests, en particulier des tests simples, est introduit au Chapitre 1 de la Partie I. Lorsque l'échantillon d'apprentissage est fini, la connaissance partielle que l'on a de P_0 et P_1 doit être exploité de manière appropriée. Le problème correspondant est le problème d'apprentissage. Celui-ci, et en particulier dans le cadre de la classification par plug-in est introduit au Chapitre 5 Partie I. Ce chapitre est aussi l'occasion de présenter la mesure d'erreur utilisée dans la Partie II : l'erreur d'apprentissage. Nous donnons quelques propriétés de celle-ci.

Dans un problème non supervisé, rien n'est donné sur les lois $(P_k)_{k=0,1}$. La structure d'hypothèses correspondante est alors une structure dans laquelle l'hypothèse nulle est simple et l'hypothèse alternative est (à peu près) tout ce qui n'est pas l'hypothèse nulle. Avec des hypothèses de ce type, on peut tester l'égalité de deux distributions ou de deux paramètres. Le Chapitre 3 de la Partie I donne une revue et une étude comparative de tests existants pour tester des hypothèses de la forme

$$H_0 \ : \ \nu = 0 \ \text{ contre } \ H_1 \ : \ \|\nu\|_{\mathbb{R}^p} \geq \rho,$$

où ν est un vecteur de $\mathbb{R}^p$ observé dans un bruit gaussien et ρ un réel positif. Un des tests exposés sera utilisé au Chapitre 3 Partie III pour la segmentation non supervisée. Nous cherchons aussi, dans la Partie I, à expliquer comment et pourquoi réduire la dimension dans la construction de procédures pour tester des hypothèses de ce type. Nous avons voulu mettre en valeur la nécessité de la réduction de dimension dans les tests d'hypothèses « trop complexes » mais ayant une dimension « effective » petite. C'est l'objet du Chapitre 2 de la Partie I.

Sur le plan pratique, une procédure de réduction de dimension peut être élaborée à partir d'une procédure de test d'hypothèses multiples. D'une certaine manière, on cherche à tester simultanément quels sont les attributs d'une observation qui sont porteur d'une information permettant de séparer des données. Ainsi, l'outil pratique essentiel utilisé pour la réduction de dimension dans ce mémoire est l'algorithme de Benjamini et Hochberg [9] pour contrôler la proportion de rejet à tord dans un problème de test d'hypothèses multiples. La problématique des tests multiples et quelques résultats à l'origine de cette théorie sont présentés dans le Chapitre

4 de la première partie.

La deuxième partie est consacrée à l'étude de la classification de données gaussiennes en grande dimension et en dimension infinie. Elle a pour objectif de présenter une méthode de réduction de dimension pour la classification, mais surtout de motiver la méthode utilisée en expliquant clairement ce qui, pour les procédures QDA et LDA, constitue un bon apprentissage.

Comme nous l'avons déjà mentionné, dans le cas de la classification binaire, un classificateur g partitionne l'espace des observations en deux régions. Dans le cas de la procédure LDA, les deux régions associées au classificateur g^* sont délimitées par un hyperplan affine H^* et celles associées au classificateur g^{LDA} sont délimitées par hyperplan affine H^{LDA}. Rappelons qu'un hyperplan affine peut être défini par deux éléments de l'espace des observations : un vecteur normal et un point que nous appelons centre. Ainsi, une méthode d'apprentissage permet d'obtenir un hyperplan H^{LDA} par un vecteur normal et un centre. Idéalement, le vecteur normal (resp. le centre) de l'hyperplan H^{LDA} doivent être proches du vecteur normal (resp. du centre) de l'hyperplan H^*. En effet, l'ensemble des éléments de l'espace des observations qui sont bien classées par la règle de Bayes et mal classés par la règle g^{LDA}, font partie de celles comprises « entre » les deux hyperplans H^{LDA} et H^*. Ainsi le vecteur normal de l'hyperplan H^{LDA} doit être construit de manière à être proche de celui de l'hyperplan H^* dans un sens bien précis, lié à la géométrie des mesures gaussiennes utilisées pour modéliser le problème. Nous quantifions de manière précise ce qui, dans une règle construite à partir d'un hyperplan, augmente l'erreur d'apprentissage. Nous montrons que le réel enjeux de l'apprentissage réside dans l'estimation du vecteur normal définissant l'hyperplan optimal et non pas dans l'estimation du centre.

Le problème d'apprentissage des paramètres définissant l'hyperplan H^* pour construire H^{LDA} (en vu de réduire l'erreur d'apprentissage) est donc très intimement lié au problème de l'estimation du vecteur normal de l'hyperplan H^*. Nous cherchons à tirer parti de ce parallèle. Dans le cadre de l'estimation d'un vecteur $\theta \in \mathbb{R}^p$ de grande dimension observé dans un bruit gaussien indépendant et identiquement distribué :

$$X_i = \theta_i + \sigma\epsilon_i, \quad i = 1, \ldots, p$$

(σ est une constante d'échelle connue) beaucoup de travaux très aboutis ont été effectués (voir par exemple l'article de revue de Candes [19]). Nous nous placerons dans le cas où le vecteur θ a peu de coefficients significatifs. Un tel vecteur est dit creux. Remarquons que si le vecteur directeur d'un hyperplan séparateur optimal H^* est creux, ceci signifie que la variabilité des données est concentrée dans un espace de petite dimension. Dans ce cas, les travaux qui répondent le mieux à nos attentes sont les travaux concernant l'estimation par seuillage. L'estimation par seuillage consiste à sélectionner à partir des attributs $(X_i)_{i\in\{1,\ldots,p\}}$ de l'observation X les composantes de θ qui sont proches de zéro et à les fixer à zéro. Le choix précis du seuil à partir duquel il est intéressant de fixer à zero une coordonnée du vecteur observé a été très largement étudié ces dix dernières années. Les travaux les plus poussés (voir l'article de Abramovich et ses collaborateurs [2]) en la matière utilisent une procédure de test multiple permettant de contrôler l'espérance de la proportion de rejet à tort (le FDR) dans le test des hypothèses :

$$H_{0i} \ : \ \theta_i = 0 \ \text{ contre } \ H_{1i} \ : \ \theta_i \neq 0 \ \ i = 1, \ldots, p. \tag{9}$$

La procédure correspondante est la procédure de Benjamini et Hochberg [9]. Dans le Chapitre 1 de la deuxième partie, nous explicitons précisément le lien entre le problème d'estimation du

vecteur normal de l'hyperplan définissant H^* et le problème d'estimation de θ. Dans le Chapitre 2 de cette même partie, nous exploitons ce lien et proposons une procédure permettant de construire un estimateur F^{LDA} du vecteur normal F^* de H^* par seuillage. Ce seuillage est effectué grâce à la méthode de contrôle du FDR de Benjamini et Hochberg. Il constitue une réduction de dimension : le classificateur g^{LDA} résultant agira dans l'espace de dimension réduite engendré par les directions dans lesquelles les coefficients du vecteur normal F^{LDA} n'ont pas été fixés à zéro.

Rappelons que, en statistique descriptive, lorsque les observations $(X_i)_{i=1,\dots,n}$ sont réparties dans K groupes différents, on peut décomposer la variance du nuage de points $(X_i)_{i=1,\dots,n}$ en une variabilité inter et une variabilité intra. La variabilité intra correspond à la variabilité des données au sein de chaque groupe et la variabilité inter correspond à la variabilité des données d'un groupe à l'autre. Une forte variabilité inter contribue à séparer les données et une faible variabilité intra aussi. Cette remarque est à l'origine de l'utilisation du rapport de Raileygh (obtenu à partir du critère de Fisher, voir par exemple [35]) en statistique descriptive multidimensionnelle. Les coefficients de F^* ont de fortes amplitudes dans les directions pour lesquelles une version théorique du rapport de variabilité inter sur la variabilité intra est grand. La procédure que nous utilisons pour choisir l'espace de dimension réduite dans lequel la règle g^{LDA} agit, a une interprétation statistique intéressante. Elle correspond à tester simultanément pour $q = 1, \dots, p$ les hypothèses

- H_{0q} : « dans la direction q, le rapport de variabilité inter sur variabilité intra est nul » ,
  contre
- H_{1q} : « Dans la direction q, le rapport de variabilité inter sur variabilité intra est non nul » .

On conserve alors les directions q dans lesquelles H_{0q} est rejetée.

Nous montrons dans le Chapitre 2 de la Partie II que l'espérance (par rapport à la loi de l'échantillon d'apprentissage) de l'erreur d'apprentissage associée au classificateur obtenu par la procédure g^{LDA} converge vers 0 à une vitesse donnée. Notons que cette convergence a lieu uniformément sur une large classe de paramètres et qu'elle reste valable lorsque la dimension p tend vers l'infini bien plus vite que la taille n de l'échantillon d'apprentissage.

Dans le cas de la procédure QDA, l'interprétation géométrique de l'ensemble des éléments de l'espace des observations mal classées par une règle g^{QDA} et bien classées par la règle optimale g^* est moins évidente. Une règle g^{QDA} donnée est définie par un centre, un vecteur normal et une matrice symétrique. Il en est de même de la règle g^*. Nous établissons le lien, dans le Chapitre 1 de la Partie II, entre l'erreur d'estimation de ces quantités associées à une règle g^{QDA} donnée et l'erreur d'apprentissage associée au classificateur g^{QDA}. Pour le vecteur normal, une procédure d'estimation similaire à celle utilisée dans le cas de la LDA est donnée. Elle donne lieu à une réduction de dimension. Nous supposons connaître une base dans laquelle la matrice associée à la règle g^* est diagonale. On peut alors tenir, pour l'estimation de la matrice symétrique définissant g^*, à peu près le même raisonnement que dans le cas de l'estimation du vecteur normal, en résumant cette matrice par le vecteur de ses valeurs propres. Nous utilisons également une procédure d'estimation par seuillage, avec un seuil défini par la procédure de Benjamini et Hochberg [9]. La procédure d'estimation correspondante a à nouveau une interprétation statistique intéressante. Elle correspond à tester simultanément pour $q \in \{1, \dots, n\}$ les hypothèses

- H_{0q} : la variabilité de la variance intra entre les différents groupes est nulle dans la direction

q,
contre
- H_{1q} : la variabilité de la variance intra entre les différents groupes est non nulle dans la direction q.

On estime la matrice associée à g^* seulement dans les directions q où H_{0q} est rejetée. Notons que cette procédure ne constitue pas une réduction de la dimension de l'espace sur lequel la fonction g^{QDA} agit. La règle g^{QDA} finalement obtenue n'est pas inactive dans les directions q pour lesquelles H_{0q} est acceptée. Dans ces directions la règle est linéaire au lieu d'être quadratique. Une règle linéaire est plus simple qu'une règle quadratique car elle fait intervenir moins de paramètres. Nous parlons donc de procédure de simplification de la règle. Remarquons que l'on pourrait parler de réduction de dimension de l'espace dans lequel se trouve g^{QDA}. Par la suite, sauf mention du contraire, nous emploierons la dénomination « réduction de dimension » pour parler de réduction de la dimension de l'espace sur lequel le classificateur g agit.

La troisième partie vise à développer des méthodes de segmentation d'images hyperspectrales. Nous avons envisagé deux méthodes différentes. Rappelons que la distribution d'une observation X_i en un pixel $i \in \mathcal{T}_N$ est donnée par l'équation (7). La première méthode de segmentation supervisée, nécessite la connaissance a priori de K et un échantillon d'apprentissage de variables aléatoires issues des lois $(P_k)_{k=1,\dots,K}$. La seconde au contraire est une méthode non supervisée, elle ne présuppose pas la connaissance de K, et ne nécessite pas d'échantillon d'apprentissage. Elle repose sur l'hypothèse que les distributions $(P_k)_k$ sont gaussiennes de covariances égales et connues à une constante d'échelle près, et que les poids π_{ik} valent 0 ou 1.

La première méthode repose sur l'hypothèse que chaque spectre est issu d'un mélange de lois normales sur un espace infini-dimensionel de moyennes et de covariances inconnues. Elle s'appuie sur une représentation multi-échelle des poids π_{ik} et une réduction de dimension par une méthode de tests multiples présentée dans la deuxième partie. L'algorithme final est une version infinie-dimensionnelle de l'algorithme de Kolaczyk et al. [46] et se décompose en trois temps. Le premier temps consiste en l'estimation des lois $(P_k)_{k=1\dots,K}$ grâce à l'échantillon d'apprentissage. Cette estimation repose sur la sélection d'un espace de dimension réduite sur lequel les observations sont projetées, et sur le choix d'un sous espace dans lequel les lois P_k ont des covariances suffisamment proches pour être supposées égales. Les densités $(P_k)_{k=1\dots,K}$ sont ensuite supposées connues. Dans le deuxième temps, les proportions du mélange définies par (7) sont estimées par maximum de vraisemblance pénalisé (c'est l'algorithme de Kolaczyk et al. [46] dans l'espace de dimension réduite). Le troisième temps consiste en la construction d'une fonction de segmentation de type plug-in à partir des proportions estimées $(\hat{\pi}_{ik})_{ik}$ et de l'estimation $(\hat{P}_k)_{k=1,\dots,K}$ des lois $(P_k)_{k=1,\dots,K}$. Cet estimateur par plug-in est celui obtenu par substitution dans (5.1) des paramètres du modèle par les paramètres estimés et par minimisation du critère résultant. D'un points de vu pratique l'originalité de la démarche réside dans l'utilisation de la première et la dernière étape de l'algorithme.

Nous étudions les performances théoriques de l'algorithme de segmentation dans le cas où $K = 2$ et $M = 2$, si les deux régions D_1 et D_2 sont séparées par une frontière qui peut être recouverte par un nombre suffisamment petit de pixels, et si les densités P_k sont supposées connues. Nous montrons que l'excès de risque de segmentation (défini par (8)) associé à notre fonction de segmentation converge vers 0 à une vitesse quasiment paramétrique. Ce résultat est basé sur

l'interprétation de l'estimateur des poids du mélange par maximum de vraisemblance pénalisé, comme un estimateur obtenu par un certain nombre de tests (cette interprétation est inspirée des travaux de Birgé sur les T-statistiques [14]). Cette interprétation n'avait pas été envisagée par Kolaczyk et al. [46].

Le deuxième algorithme peut être assimilé à un croisement entre un algorithme de segmentation par croissance de régions et un algorithme de lissage par noyau. La méthode correspondante étend la méthode AWS proposée par Polzehl et Spokoiny [59] pour le débruitage d'images multidimensionnelles. Nous supposons observer sur chaque pixel i de l'image,

$$X_i = \mu_i + \sigma\epsilon_i \quad \text{où} \quad \mu_i = \sum_{m=1}^{M} a_m 1_{i\in D_m},$$

$(a_m)_{m=1,\dots,M}$ sont des courbes inconnues toutes différentes, et $(D_m)_m$ sont des régions inconnues formant une partition de l'image. La méthode AWS est une méthode adaptative de lissage à noyau. La fenêtre de lissage du noyau est sélectionnée de manière récursive par des procédures de tests d'hypothèses qui pour $i, j \in \mathcal{T}_N$ sont du type

$$H_0 \ : \mu_i = \mu_j \quad \text{contre} \quad H_1 \ \ \mu_i \neq \mu_j. \tag{10}$$

Le cas d'images hyperspectrales n'est pas celui étudié par Polzehl et Spokoiny [59], les paramètres inconnus μ_i sont infini-dimensionnels. Nous étendons donc l'algorithme de lissage d'images, AWS, au cadre des images hyperspectrales en utilisant des tests fonctionnels. Nous introduisons aussi dans l'algorithme la procédure de test des hypothèses multiples

$$H_{0j} \ : \mu_i = \mu_j \ \ \text{contre} \ \ H_{1j} \ : \mu_j \neq \mu_i \ \ j \in V_i,$$

où V_i est un ensemble de pixels voisins du pixel i.

L'algorithme AWS ne fournit pas une segmentation de l'image mais les poids du noyau donnent une mesure de similarité entre les pixels. L'algorithme fournit des poids w_{ij} valant 0 ou 1. Ces poids imitent une règle parfaite qui saurait décider laquelle des hypothèses définies par (10) est vraie. Nous utilisons le fait que, dans notre image, deux pixels i et j sont séparés par une frontière si et seulement si pour tout pixel q de l'image

$$(\mu_i = \mu_q \quad \text{et} \quad \mu_j \neq \mu_q) \quad \text{ou} \quad (\mu_i \neq \mu_q \quad \text{et} \quad \mu_j = \mu_q).$$

Cette caractérisation des frontières nous conduit à construire une statistique à partir de

$$TV(i,j) = \sum_p Vote(q)(i,j),$$

où $Vote(q)(i,j) = 1_{w_{iq}\neq w_{jq}}$ caractérise le fait qu'un pixel q vote pour que les deux pixels adjacents i et j soient séparés par une frontière et $TV(i,j)$ est le total des votes pour que cette frontière existe. La statistique ainsi définie nous permet d'estimer les frontières séparant les différentes zones et d'en déduire une segmentation de l'image. L'algorithme est illustré sur les images de l'INSERM et sur une image test construite pour ce type de problème. Les propriétés de l'algorithme de lissage AWS ne sont établies (sous des hypothèses simplificatrices) que dans le cas d'une image lisse. Nous ne sommes pas parvenus à obtenir de résultats théoriques dans le cadre de la segmentation, mais ceux obtenus dans la pratique sont très convaincants. Ces résultats pratiques et la description de l'algorithme que nous proposons ont été publiés dans un numéro spécial de « la revue traitement du signal » consacré à la cancérologie [36].

Première partie

Classification et tests d'hypothèses

> Ceux qui, dans le formidable espace de l'esprit humain, semblent apporter des éléments nouveaux et parler une nouvelle langue, ne font que traduire au travers de leur sensibilité propre ce que les autres ont déjà pensé et dit.
>
> Gabriel Fauré

> Don't just read it ; fight it ! Ask your own questions, look for your own examples, discover your own proofs. Is the hypothesis necessary ? Is the converse true ? What happens in the classical special case ? What about the degenerate cases ? Where does the proof use the hypothesis ?
>
> Paul Halmos, "I want to be a mathematician".

Dans cette première partie nous rappelons les bases à la fois des tests d'hypothèses (théorie de la décision statistique) et de la classification. Les deux problèmes sont en de nombreux points liés. Dans les deux cas (si l'on se restreint à la classification binaire) il s'agit de choisir entre deux alternatives tout en contrôlant l'erreur consistant à faire un mauvais choix. Cet objectif est au coeur de notre problématique. Cette partie est composée de 5 chapitres, liés entre eux par la problématique des tests. La plupart des thèmes traités sont orientés vers les problèmes posés par la grande dimension et vers les techniques de réduction de dimension. Cette partie ne comprend pas de résultats singulièrement nouveaux (en dehors du Théorème 5.1 du Chapitre 5) mais quelques formulations éclairants la problématique de la grande dimension ainsi que quelques introductions à des outils utilisés par la suite.

Dans le premier chapitre, nous décrivons les tests d'hypothèses dans un cadre assez général. Nous faisons le tour de la question du test simple pour souligner le lien entre erreur de test et norme L_1 entre deux mesures. Nous étudions dans ce chapitre le problème de détection qui sert de fondation à la deuxième partie de ce mémoire.

Les Chapitres 2 et 3 sont eux motivés par la dernière partie de ce mémoire. Le Chapitre 2 introduit la problématique des tests minimax. Quelques démonstrations de résultats existant sont données afin d'illustrer simplement la nécessité de séparer les hypothèses statistiques et, dans certains cas, de réduire la dimension. Le Chapitre 3 introduit les tests d'hypothèses sur la norme d'un vecteur gaussien de $\mathbb{R}^p$. Dans ces hypothèses, l'hypothèse nulle correspond à la nullité d'un paramètre, et l'hypothèse alternative correspond à la non nullité de ce paramètre. Nous expliquons en quoi ce type de spécification des hypothèses est lié à la classification et la segmentation non supervisées. La taille de l'ensemble des hypothèses alternatives (qui constitue une méconnaissance de l'hypothèse alternative) est à l'origine d'une nécessaire réduction de dimension. Le Chapitre 4 concerne les tests d'hypothèses multiples. L'algorithme de Benjamini et Hochberg y est présenté. Des variantes de celui-ci sont utilisées dans la deuxième partie pour la réduction de dimension et dans la troisième partie pour la segmentation non supervisée.

Dans le Chapitre 5 de cette partie, nous introduisons le problème de classification supervisée

et plus particulièrement le problème de learning par plugin. Le problème de classification supervisée sera central dans les deux dernières parties de ce mémoire.

Chapitre 1

Généralités sur les tests

> Il n'y a pas de vérité première, il n'y a que des erreurs premières.
>
> Bachelard

Dans ce chapitre, nous introduisons les tests d'hypothèses statistiques. Nous décrivons les tests optimaux dans le cadre d'hypothèses simples. Nous donnons l'exemple du problème de détection. Nous donnons l'importance, dans le cadre des tests, de la distance L_1 entre deux mesures et son lien avec d'autres distances statistiques.

1.1 Quelques notations

Avant de parler de tests, nous introduisons des notations et définitions (liées aux tests) qui seront utilisées par la suite.

Si $(\mathcal{X}, \mathcal{A})$ est un espace mesurable, nous dirons qu'une mesure positive μ σ-finie est **absolument continue** par rapport à une mesure ν (notée $\mu \ll \nu$), si pour tout $A \in \mathcal{A}$, $\nu(A) = 0$ implique $\mu(A) = 0$. Dans toute la suite, les mesures considérées seront positives et σ-finies. Une mesure μ est dite **équivalente** à ν (noté $\mu \sim \nu$), si $\mu \ll \nu$ et $\nu \ll \mu$. Nous appellerons **support** d'une mesure μ sur $(\mathcal{X}, \mathcal{A})$ un espace topologique mesuré, la fermeture de l'ensemble des $x \in X$ tels que $\mu(N_x) > 0$ pour tout ouvert N_x contenant x.

Rappelons que si $\mu \ll \nu$, il existe une application $\mathcal{A}$-mesurable de $\mathcal{X}$ dans $\mathbb{R}^+$ déterminée de manière unique pour ν-presque tout $x \in \mathcal{X}$ telle que pour tout $A \in \mathcal{A}$, $\mu(A) = \int_A d\nu$. Une telle fonction sera dite **dérivée de Radon-Nikodym** de μ par rapport à ν et notée $\frac{d\mu}{d\nu}$. Un ensemble de mesures de probabilité $\mathcal{P}$ est dit **dominé** par une mesure λ si toute mesure P de $\mathcal{P}$ est absolument continue par rapport à la mesure λ. Par exemple, deux mesures finies (ν, μ) sont dominées par la mesure finie $\eta = \lambda + \mu$. Aussi, pour deux mesures finies, nous écrirons $\{d\mu > \alpha d\nu\}$ $(\alpha \in \mathbb{R})$ à la place de $\{x \in \mathcal{X} \ : \ \frac{d\mu}{d\eta}(x) > \frac{d\nu}{d\eta}(x)\alpha\}$, ou encore $\int |d\nu - d\mu|$ pour $\int |\frac{d\nu}{d\eta} - \frac{d\mu}{d\eta}| d\eta$.

Deux mesures μ et ν seront dites **orthogonales** s'il existe $A \in \mathcal{A}$ tel que $\nu(A) = \mu(\mathcal{X} \setminus A) = 0$, ce qui sera noté $\nu \perp \mu$.

Nous noterons $\mathbb{E}_P$ l'espérance calculée par rapport à la loi P. Lorsque qu'il n'y aura pas d'ambiguïté sur la nature de la loi utilisée, nous noterons $\mathbb{E}_n$ au lieu de $\mathbb{E}_{P_n}$ (pour n un entier positif et P_n une loi de probabilité) et $\mathbb{E}$ au lieu de $\mathbb{E}_P$.

Par la suite, si p est un entier positif, nous noterons $\langle ., . \rangle_{\mathbb{R}^p}$ le produit scalaire usuel de $\mathbb{R}^p$ et $\|\|_{\mathbb{R}^p}$ la norme associée.

1.2 Généralités

1.2.1 Problématique, définition et exemples

Problématique des tests. Soit $\mathcal{P}$ l'ensemble des mesures de probabilité sur $(\mathcal{X}, \mathcal{A})$ un espace mesurable. La problématique des tests d'hypothèses est la suivante. Etant donnés deux sous-ensembles disjoints (non vides) $\mathcal{P}_0$ et $\mathcal{P}_1$ de $\mathcal{P}$ et une (ou plusieurs) observation(s) $X \in \mathcal{X}$ issue(s) d'une loi inconnue $P \in \mathcal{P}_0 \cup \mathcal{P}_1$ on veut décider si $P \in \mathcal{P}_0$ ou si $P \in \mathcal{P}_1$.

Les ensembles $\mathcal{P}_0$ et $\mathcal{P}_1$ sont appelés respectivement ensemble d'hypothèses nulles et ensemble d'hypothèses alternatives. La décision « ne pas rejeter l'hypothèse nulle » correspond à décider que $P \in \mathcal{P}_0$ et « rejeter l'hypothèse nulle » correspond à décider que $P \in \mathcal{P}_1$.

Remarque 1.1 (terminologie)**.** *Dans le premier cas on dit aussi que l'on accepte l'hypothèse nulle. Il faut noter que dans les domaines d'application des tests cette deuxième dénomination amène parfois à des interprétations erronées et l'expression « ne pas rejeter » est donc préférable. Gardant à l'esprit que prendre une décision ne s'apparente pas pour nous à établir une vérité, nous utiliserons indifféremment les deux terminologies.*

Lorsqu'un ensemble d'hypothèse est réduit à un élément, on parle d'hypothèse simple.

Définition des tests. Un test est une fonction qui permet de prendre la décision quant à l'appartenance de la loi inconnue P à l'ensemble $\mathcal{P}_0$ ou à $\mathcal{P}_1$. Concrètement nous appellerons test non randomisé toute application (mesurable) $\psi : \mathcal{X} \to \{0, 1\}$. La valeur 0 correspond à accepter H_0 (i.e. : décider que $P \in \mathcal{P}_0$) et la valeur 1 correspond à rejeter H_0. D'un point de vue pratique ce type d'application recouvre tout ce qui peut être envisagé pour décrire une décision binaire : être ou ne pas être dans $\mathcal{P}_0$, mais pour des raisons mathématiques il est nécessaire d'utiliser une classe de fonctions de décision plus grande.

Définition 1.1. *Un test est un élément de l'ensemble des tests :*

$$\Psi = \{\psi : \mathcal{X} \to [0, 1] \text{ } \textit{mesurables} \text{ }\}.$$

Si pour $x \in \mathcal{X}$ la fonction $\psi \in \Psi$ vaut $\psi(x) = p \in [0, 1]$ on accepte H_0 de manière aléatoire, avec une probabilité $1 - p$. Si pour tout x, $\psi(x)$ vaut 0 ou 1 le test est non-randomisé.

Tous les tests considérés par la suite seront randomisés. Dans la pratique, utiliser un test randomisé permet de ne pas avoir de parti-pris pour l'une ou l'autre des hypothèses, quand celles-ci ne sont pas vraiment distinguables (ce terme sera défini ultérieurement). En théorie, si un test optimal est randomisé, sur un ensemble de mesure non nulle, cela signifie que les hypothèses à tester ne sont pas distinguables sur cet ensemble (i.e. sur cet ensemble, il est préférable de prendre une décision au hasard).

Propriétés de Ψ. Tout d'abord, $\Psi \subset \{\text{fonctions mesurables bornées sur } \mathcal{X}\}$. La propriété mathématique essentielle de Ψ est la convexité. Dans le cas où les ensembles de mesures $\mathcal{P}_0$ et $\mathcal{P}_1$ sont dominés par une mesure λ, Ψ est un sous-ensemble fermé borné de $L^\infty(\lambda)$, il est donc compact pour la topologie faible$-*$ de $L^\infty(\lambda)$ (d'après le théorème de Banach-Alaoglu, cf Théorème p66 de [65]). Nous ne donnons pas ici la définition de la faible$-*$ compacité mais si λ est σ-finie alors (cf Théorème p158 de [64]) $L^\infty(\lambda) = L^1(\lambda)^*$, et la faible-* compacité signifie que si ψ_n est une suite de tests, on peut en extraire une sous-suite ψ_{n_k} convergeant vers un test ψ au sens suivant :

$$\forall u \in L^1(\lambda) \quad \lim_{k\to\infty} \int \psi_{n_k}(x)u(x)d\lambda(x) = \int \psi(x)u(x)d\lambda(x).$$

On en déduit (en prenant $u = \frac{dP}{d\lambda}$) le théorème suivant :

Theoreme 1.1. *Soit $\mathcal{P}$ un ensemble de mesures positives dominées par une mesure λ σ-finie. L'application bilinéaire continue L qui à $(P, \psi) \in \mathcal{P} \times \Psi$ associe $\mathbb{E}_P[\psi]$ a une image fermée. En particulier l'application qui a $\psi \in \Psi$ associe $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\psi]$ a une image fermée et pour toute mesure de probabilité P l'application qui à $\psi \in \Psi$ associe $\mathbb{E}_P[\psi]$ a une image fermée.*

Ce théorème quelque peu abstrait en apparence assurera l'existence de tests optimaux.

Comparaison des tests. La meilleure manière de mesurer la qualité d'une décision liée à des phénomènes aléatoires est de mesurer la probabilité de prendre une mauvaise décision. Si $P \in \mathcal{P}_0$, l'erreur correspondante est appelée erreur de première espèce et notée

$$\alpha(\psi, P) = \mathbb{E}_P[\psi]. \tag{1.1}$$

Si $P \in \mathcal{P}_1$, l'erreur correspondante est appelée erreur de seconde espèce et notée

$$\beta(\psi, P) = 1 - \mathbb{E}_P[\psi]. \tag{1.2}$$

Dans la suite, lorsqu'il n'y aura pas d'ambiguïté, nous noterons $\beta(\psi)$ au lieu de $\beta(\psi, P)$ et $\alpha(\psi)$ au lieu de $\alpha(\psi, P)$. Dans notre formalisme, la probabilité de faire une erreur est donc :

$$\mathcal{E}(\psi, P) = \begin{cases} \alpha(\psi, P) & \text{si } P \in \mathcal{P}_0 \\ \beta(\psi, P) & \text{si } P \in \mathcal{P}_1 \end{cases}.$$

Elle est fonction de la loi inconnue P et du test ψ.

Nous dirons qu'il existe un test parfait $\psi^* \in \Psi$, si ψ^* permet d'obtenir une erreur minimale uniformément en P, i.e si

$$\forall(\psi, P) \in \Psi \times (\mathcal{P}_0 \cup \mathcal{P}_1) \quad \mathcal{E}(\psi^*, P) \leq \mathcal{E}(\psi, P).$$

Remarquons que l'existence d'un test parfait est déterminée par la proposition suivante.

Proposition 1.1. *Dans le cas d'hypothèses simples $\{P_0\}$ et $\{P_1\}$, il existe un test parfait si et seulement si $P_1 \perp P_0$. Dans le cas d'hypothèses composées d'un nombre dénombrable d'éléments une condition nécessaire et suffisante à l'existence du meilleur test est*

$$\forall(P_0, P_1) \in \mathcal{P}_0 \times \mathcal{P}_1 \quad P_0 \perp P_1. \tag{1.3}$$

Démonstration. Si $P_1 \perp P_0$ alors il existe $A \in \mathcal{A}$ tel que $P_1(A) = P_0(\mathcal{X} \setminus A) = 0$. Ainsi, pour le test

$$\psi^* = 1_{\mathcal{X} \setminus A}, \tag{1.4}$$

on obtient $\mathcal{E}(\psi^*, P_1) = \mathcal{E}(\psi^*, P_0) = 0$ ce qui est optimal. S'il existe un meilleur test ψ^*, on a $\mathbb{E}_1[1 - \psi^*] = 0$ et $\mathbb{E}_0[\psi^*] = 0$ et donc en notant $A^* = \{x \in \mathcal{X} \ : \ \psi^*(x) > 0\}$, $B^* = \{x \in \mathcal{X} \ : \ \psi^*(x) < 1\}$ on a

$$P_1(B^*) = 0 \ \text{ et } P_0(A^*) = 0.$$

On conclut alors, pour le cas des hypothèses simples, en notant que $\mathcal{X} \setminus A^* \subset B^*$ et que donc $P_1(\mathcal{X} \setminus A^*) = 0$. Pour les hypothèses composées de plusieurs éléments si (1.3) est vérifiée, alors, pour tout $P_0, P_1 \in \mathcal{P}_0, \mathcal{P}_1$, il existe $A_{P_0,P_1} \in \mathcal{A}$ tel que $P_1(A_{P_0,P_1}) = P_0(\mathcal{X} \setminus A_{P_0,P_1}) = 0$. En prenant $A = \cap_{P_0 \in \mathcal{P}_0} \cup_{P_1 \in \mathcal{P}_1} A_{P_0,P_1}$, on a bien $P_0(A) \leq P_0(\cup_{P_1 \in \mathcal{P}_1} A_{P_0,P_1}) = 0$ et $P_1(\mathcal{X} \setminus A) \leq P_1(\cup_{P_0 \in \mathcal{P}_0}(\mathcal{X} \setminus A_{P_0,P_1})) = 0$. Réciproquement s'il existe un meilleur test ψ^*, ce test est aussi un meilleur test pour tester toutes les hypothèses simples obtenues en prenant une hypothèse dans $\mathcal{P}_0$ et une dans $\mathcal{P}_1$, ce qui implique (1.3). □

D'un point de vue statistique l'orthogonalité joue donc le rôle de modèle « parfait » . Le meilleur test est donné par (1.4).

1.3 Etude du test d'hypothèses simples

Ensemble Δ des erreurs possibles. Etant donnée l'impossibilité (hormis dans des cas pathologiques) de trouver un test parfait dans le cas d'hypothèses simples, il est intéressant de chercher à décrire conjointement les erreurs faites sous H_0 et sous H_1 en fonction du test choisi. Pour cela nous allons étudier l'ensemble de $[0,1]^2$ suivant :

$$\Delta = \left\{(\alpha(\psi), \beta(\psi)) \in [0,1]^2 \ \ tq \ \ \psi \in \Psi\right\}.$$

Le Théorème 1.1 nous permet d'affirmer que Δ est fermé, et donc compact ; il est convexe[1], en remarquant que si $\psi \in \Psi$ alors $1 - \psi \in \Psi$ on voit que $(1/2, 1/2)$ est son centre de symétrie.

D'autre part, le test qui correspond à rejeter tout le temps H_0 et celui qui correspond à accepter tout le temps H_0 nous montrent que les points $(1,0)$ et $(0,1)$ sont dans Δ. Soit ψ_0 un test qui a la plus petite erreur de seconde espèce a parmi les tests ayant une erreur de première espèce nulle ($\alpha(\psi_0) = 0$). Soit ψ_1 un test qui a la plus petite erreur de première espèce b parmi les tests ayant une erreur de seconde espèce nulle ($\beta(\psi_1) = 0$). Notons que ces deux tests existent en vertu du Théorème 1.1. Les points $A = (0, a)$ et $B = (b, 0)$ sont des points du bord de Δ. Si un de ces deux points se trouve sur $(0,0)$ alors les deux le sont (par convexité). Il existe alors un test parfait et les probabilités P_0 et P_1 sont orthogonales. La réciproque est vraie. Supposons que les probabilités ne sont pas orthogonales, nous allons décrire la frontière qui relie le point A au point B, au moyen de deux fonctions différentes.

1. Cette frontière est le graphe de la fonction définie par :

$$g_1(\alpha_0, P_1, P_0) = \inf_{\alpha(\psi) \leq \alpha_0} \beta(\psi). \tag{1.5}$$

[1] C'est l'image du convexe $\Psi \times \Psi$ par la composition de l'application linéaire $(\mathbb{E}_P[.], \mathbb{E}_P[.])$ et de l'application qui à $(x,y) \in [0,1]^2$ associe $(x, 1-y)$ toutes deux préservant la convexité

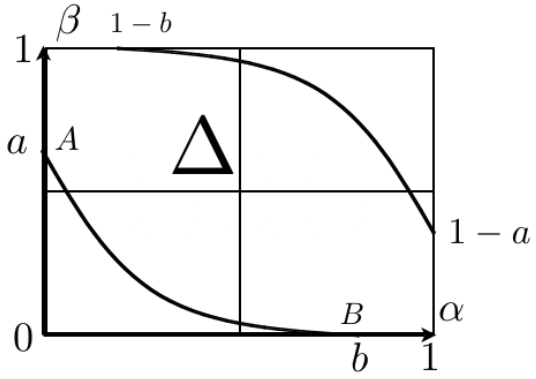


FIG. 1.1 – Le domaine d'erreur possible

2. Plutôt que de fixer l'erreur de première espèce (l'abscisse) et de regarder l'erreur de seconde espèce (l'ordonnée), il est souvent plus adapté de chercher à contrôler une combinaison convexe des deux erreurs. Nous noterons donc

$$g_2(t, P_1, P_0) = \inf_{\psi}(t\alpha(\psi, P_0) + (1-t)\beta(\psi, P_1)). \tag{1.6}$$

Nous allons montrer l'existence de ψ_t^* réalisant l'infimum ci-dessus, la frontière est alors la courbe paramètrée $(\alpha(\psi_t^*), \beta(\psi_t^*))$ quand t parcourt $[0,1]$.

Puisque Δ est convexe, g est convexe et puisque le segment $[AB]$ a une pente négative strictement, g_1 est décroissante et a une dérivée négative strictement à gauche et à droite sur $[0, b[$. La fonction g_1 est continue sur $]0, b[$ car elle est convexe sur $[0,1]$, elle est aussi continue en A ($\alpha = 0$) et B ($\alpha = b$) car son graphe est une partie connexe du bord de Δ fermé, elle admet pour cette même raison une dérivée à gauche et à droite partout, et ces dérivés se recollent sauf en un nombre dénombrable de points. Toutes les informations qui précèdent sont résumées sur la figure (1.1).

Nous allons maintenant identifier les tests optimaux qui permettent d'atteindre les points de la partie du bord de Δ définie par le graph de g_1 sur $[0, b]$.

Description des tests optimaux. La frontière g_1 (définie par (1.5)) correspond à la solution d'un problème de minimisation de la fonction convexe $\beta(\psi)$ (1 − une fonction linéaire) sous la contrainte linéaire que $\alpha(\psi) - \alpha_0$ soit plus petit ou égal à 0, aussi on peut facilement établir

$$\inf_{\psi\,:\,\alpha(\psi)\le\alpha_0} \beta(\psi) = \inf_{\psi\in\Psi} \sup_{t\in[0,\infty]} \mathcal{L}(t,\psi), \tag{1.7}$$

où $\mathcal{L}(t,\psi)$ est le Lagrangien du problème d'optimisation considéré :

$$\mathcal{L}(t,\psi) = \beta(\psi) + t(\alpha(\psi) - \alpha_0) = 1 - t\alpha_0 - \int \psi(dP_1 - t dP_0). \tag{1.8}$$

La solution $\psi^*(\alpha_0)$ de ce problème est garantie par le Théorème (1.1), et on peut donc ici appliquer un théorème minimax[2] (cf [33]) pour obtenir

$$\inf_{\psi\,:\,\alpha(\psi)\le\alpha_0} \beta(\psi) = \sup_{t\in[0,\infty]} \inf_{\psi\in\Psi} \mathcal{L}(t,\psi) = \mathcal{L}(t_0, \psi^*(t_0)),$$

[2]Nous démontrons dans le Chapitre 3 de cette partie le lemme minimax duquel le théorème de Fan découle par un argument topologique.

où au vu du dernier membre de l'équation (1.8) :

$$\psi^*(t) = \text{Arginf}_{\psi\in\Psi}\, 1 - t\alpha_0 - \int \psi(dP_1 - tdP_0) = \left\{ \begin{array}{cc} 1 & \text{si } dP_1 > tdP_0 \\ c(t) \in [0,1] & \text{si } dP_1 = tdP_0 \\ 0 & \text{si } dP_1 < tdP_0 \end{array} \right. , \qquad (1.9)$$

et

$$t_0 = \text{Argsup}_{t\in[0,\infty], c(t)\in[0,1]}\, \mathcal{L}(t, \psi^*(t)) = \sup\left\{t \in [0,\infty]\ c(t) \in [0,1]\ :\ \alpha(\psi^*(t)) \leq \alpha_0\right\}.$$

La famille de tests $\psi^*(t_0)$ quand α_0 varie est la famille des tests de Neymann-Pearson. Notons que le seuil t_0 n'est pas parfaitement déterminé. La recherche de t_0 est bien la partie la plus délicate de la construction du test, il n'existe pas de méthode générale pour l'obtenir. Il existe cependant des méthodes pour obtenir une valeur inférieure assez bonne de t_0. Nous reviendrons sur ce problème dans la suite. Remarquons enfin que si

$$P(\{A \in \mathcal{A}\ :\ P_1(A) = t_0 P_0(A)\}) = \int 1_{\{dP_1 = t_0 dP_0\}}(x) dP = 0,$$

alors le test de Neymann-Pearson est non randomisé. Dans le cas où $\sup_{A\in\mathcal{A}} |P_1(A) - P_0(A)| = 0$ le test est totalement randomisé, cela signifie que les lois P_1 et P_0 sont indistiguables par une règle de décision (dans toutes les situations, la meilleur stratégie est de choisir une des deux hypothèses au hasard).

Dans le cas où l'on cherche à contrôler $t\alpha(\psi) + (1-t)\beta(\psi)$, on montre facilement que par le même type d'écriture que dans (1.9), on a :

$$\text{Arginf}_{\psi\in\Psi}\, (t\alpha(\psi, P_0) + (1-t)\beta(\psi, P_1)) = \psi^*(t) = \left\{ \begin{array}{cc} 1 & \text{si } dP_1 > \frac{t}{1-t} dP_0 \\ c(t) & \text{si } dP_1 = \frac{t}{1-t} dP_0 \\ 0 & \text{si } dP_1 < \frac{t}{1-t} dP_0 \end{array} \right. , \qquad (1.10)$$

$$\text{où } c(t) = \left\{ \begin{array}{cc} 1 & \text{si } t > 1/2 \\ c \in [0,1] & \text{si } t = 1/2 \\ 0 & \text{si } t < 1/2 \end{array} \right. .$$

Cette règle de décision correspond, en classification dans le paradigme bayesien à la règle de Bayes. En effet si t est la probabilité a priori que l'hypothèse H_0 soit vraie et que P_i est la loi de probabilité du phénomène observé conditionnellement au fait que H_i $(i = 0, 1)$ soit vraie, alors la règle de Bayes est la règle de décision ψ^* qui minimise la probabilité de prendre une mauvaise décision qui est d'après la formule de Bayes :

$$P(\psi^* = 0 | H_0 \text{ vrai }) P(H_0 \text{ vrai }) + P(\psi^* = 0 | H_1 \text{ vrai }) P(H_1 \text{ vrai }) = t\alpha(\psi) + (1-t)\beta(\psi).$$

1.4 Un exemple : le problème de détection.

Avant d'étudier le problème de détection, nous allons introduire quelques notations que nous utiliserons dans le reste du mémoire. Si A est une matrice symétrique sur $\mathbb{R}^p$, nous noterons

$$q_A(x) = \langle Ax, x\rangle_{\mathbb{R}^p}. \qquad (1.11)$$

Notons que l'application qui à $x \in \mathbb{R}^p$ associe q_A est toujours la différence de deux semi-normes au carré et n'est le carré d'une norme que si A est définie positive. On a tout de même pour cette fonction les identités suivantes :

$$\forall x, y \in \mathbb{R}^p \quad q_A(x) + q_A(y) = \frac{1}{2}\left(q_A(x+y) + q_A(x-y)\right) \text{ Identité du parallélogramme} \tag{1.12}$$

$$\forall x, y \in \mathbb{R}^p \quad \langle Ax, y\rangle_{\mathbb{R}^p} = \frac{1}{4}\left(q_A(x+y) - q_A(x-y)\right) \text{ Identité de polarisation .} \tag{1.13}$$

On appelle problème de détection gaussienne le problème de test d'hypothèses simples gaussiennes, autrement dit, le problème dans lequel P_1 et P_0 sont deux lois normales sur $\mathcal{X} = \mathbb{R}^p$, γ_{C_1,μ_1} et γ_{C_0,μ_0} de moyennes respectives μ_1, μ_0 et de covariances C_1, C_0. Ce problème jouera un rôle important dans la suite de ce mémoire. Dans toute la suite du mémoire, nous allons noter

$$\mathcal{L}_{10}(x) = \ln\left(\frac{dP_1}{dP_0}(x)\right) \tag{1.14}$$

(Le Lagrandien $\mathcal{L}(t, \psi)$ défini par (1.8) n'a rien à voir avec cette notation et ne sera utilisé que dans ce Chapitre).

Proposition 1.2. *Dans le cas où $C_1 \neq C_0$, $\mathcal{L}_{10} = \mathcal{L}^Q_{10}(x)$ est un polynôme de degré deux sur $\mathbb{R}^p$.*

$$\mathcal{L}^Q_{10}(x) = -\frac{1}{2}q_{A_{10}}(x - s_{10}) + \langle G_{10}, x - s_{10}\rangle_{\mathbb{R}^p} - c \tag{1.15}$$

où

$$A_{10} = C_1^{-1} - C_0^{-1}, \;\; G_{10} = S_{10}m_{10}, \;\; S_{10} = \frac{C_0^{-1} + C_1^{-1}}{2}, \;\; c = \frac{1}{8}q_{A_{10}}(m_{10}) + \frac{1}{2}\log|\det(C_0^{-1}C_1)|,$$

m_{10} et s_{10} sont définis par

$$m_{10} = \mu_1 - \mu_0 \text{ et } s_{10} = \frac{\mu_1 + \mu_0}{2}.$$

Démonstration. Le logarithme du rapport de vraisemblance de γ_{C_1,μ_1} par rapport à γ_{C_0,μ_0} est donné par

$$\mathcal{L}^Q_{10}(x) = \frac{1}{2}q_{C_0^{-1}}(x - \mu_0) - \frac{1}{2}q_{C_1^{-1}}(x - \mu_1) + \frac{1}{2}\log|\det(C_1^{-1}C_0)|. \tag{1.16}$$

Notons que $x - \mu_1 = x - s_{10} - \frac{1}{2}m_{10}$ et $x - \mu_0 = x - s_{10} + \frac{1}{2}m_{10}$. Ainsi, d'une part l'identité de polarisation (1.13) implique que

$$q_{C_0^{-1}}(x - \mu_0) - q_{C_0^{-1}}(x - \mu_1) = 2\langle C_0^{-1/2}m_{10}, x - s_{10}\rangle_{\mathbb{R}^p}, \tag{1.17}$$

et d'autre part l'identité du parallélogramme (1.12) implique que

$$1/2\left(q_{A_{10}}(x - \mu_1) + q_{A_{10}}(x - \mu_0)\right) = q_{A_{10}}(x - s_{10}) + \frac{1}{4}q_{A_{10}}(m_{10}). \tag{1.18}$$

Notons que

$$q_{C_0^{-1}}(x - \mu_0) - q_{C_1^{-1}}(x - \mu_1) = -q_{A_{10}}(x - \mu_1) + q_{C_0^{-1}}(x - \mu_0) - q_{C_0^{-1}}(x - \mu_1), \tag{1.19}$$

D'après les équation (1.19) et (1.16) et le fait que $\mathcal{L}_{10}^Q = -\mathcal{L}_{01}^Q$, $A_{10} = -A_{01}$, on a

$$\begin{aligned}4\mathcal{L}_{10}^Q &= -(q_{A_{10}}(x-\mu_1) + q_{A_{10}}(x-\mu_0)) \\ &\quad + q_{C_0^{-1}}(x-\mu_0) - q_{C_0^{-1}}(x-\mu_1) - q_{C_1^{-1}}(x-\mu_1) + q_{C_1^{-1}}(x-\mu_0) \\ &= -2q_{A_{10}}(x-s_{10}) - \frac{1}{2}q_{A_{10}}(m_{10}) + 4\langle S_{10}m_{10}, x-s_{10}\rangle_{\mathbb{R}^p},\end{aligned}$$

où la dernière égalité résulte de l'application de l'identité de polarisation (1.17) et de l'identité du parallélogramme (1.18). □

Notons que le rapport de vraisemblance s'écrit aussi :

$$\frac{d\gamma_{C_1,\mu_1}}{d\gamma_{C_0,\mu_0}} = e^{-D_1(x)+D_0(x)} \text{ où } D_i(x) = \frac{1}{2}\left(q_{C_i^{-1}}(x-\mu_i) + \log(\det(C_i))\right),\ \ i=0,1.$$

est la distance de Mahalanobis (liée à la norme auto reproduisante que nous introduirons dans la deuxième partie de ce mémoire). Nous nous intéresserons aussi au problème infini-dimensionel dans lequel $\mathcal{X}$ est un espace de Banach. L'utilisation d'une telle abstraction sera motivée dans la partie II et le formalisme correspondant est introduit en annexe. Notons seulement que deux mesures gaussiennes sur un espace de Banach sont soit orthogonales soit équivalentes et que les conditions nécessaires et suffisantes pour l'orthogonalité sont parfaitement connues. Lorsque les mesures P_0 et P_1 à tester sont orthogonales, il existe un meilleur test donné par la connaissance d'un borélien A de $\mathcal{X}$ tel que $P_0(A) = 1$ et $P_1(A) = 0$. Dans le cas contraire les mesures sont équivalentes et alors il faut choisir un critère d'erreur donnant lieu à la règle de Bayes ou de Neyman-Pearson (selon le problème considéré). Nous étudierons ce type de problème de test lorsque les paramètres modélisant les deux hypothèses sont mal spécifiés.

Dans le cas gaussien, on peut aussi effectuer le calcul de la distance L_1. Nous aurons besoin du résultat suivant dans la suite.

Proposition 1.3. *1. Soient P_1, P_0 deux mesures de probabilité, P une mesure de probabilité qui domine P_1 et P_0,*

$$f_{10}(P,X) = \frac{1}{2}\log\left(\frac{dP_1 dP_0}{dP^2}\right), \tag{1.20}$$

et $g : \mathcal{X} \to \mathbb{R}$ mesurable et intégrable par rapport à P_0 et P_1. On a :

$$\begin{aligned}\int_{\mathcal{X}} g(x)|dP_1 - dP_0| &= 2\mathbb{E}_P\left[g(X)e^{f_{10}(P,X)}|\sinh(\mathcal{L}_{10}(X)/2)|\right], \\ \int_{\mathcal{X}} g(x)(dP_1 + dP_0) &= 2\mathbb{E}_P\left[g(X)e^{f_{10}(P,X)}|\cosh(\mathcal{L}_{10}(X)/2)|\right].\end{aligned}$$

2. Supposons que pour $i = 0, 1$, P_i est une mesure gaussienne sur $\mathbb{R}^p$ de moyenne μ_i et de covariance C_i, que $P_{1/2}$ est la mesure gaussienne de moyenne s_{10} et de covariance S_{10}^{-1}, que $X \rightsquigarrow P_{1/2}$ et $U \rightsquigarrow \gamma_{I_p,0}$. On a alors en loi les égalité suivante

$$f_{10}(P_{1/2},X) = \langle S_{10}^{-1/2}A_{10}, U\rangle_{\mathbb{R}^p} - \frac{1}{4}\log\left(|det(S_{10}^2 C_0 C_1)|\right) - \frac{1}{8}q_{S_{10}}(m_{10}), \tag{1.21}$$

$$\mathcal{L}_{10}(X) = -\frac{1}{2}q_{S_{10}^{-1/2}A_{10}S_{10}^{-1/2}}(U) + \langle S_{10}^{1/2}m_{10}, U\rangle_{\mathbb{R}^p} - \frac{1}{2}\log(|\det(C_0^{-1}C_1)|). \tag{1.22}$$

3. De plus, si $C_1 = C_0 = C$, alors

$$d_1(P_1, P_0) = 1 - 2\Phi\left(\frac{1}{2}\|C^{-1/2}m_{10}\|_{\mathbb{R}^p}\right)$$

Démonstration. Le point 1 de la proposition est immédiat, il suffit de remplacer $f_{10}(P, X)$ par sa valeur et de développer le membre de droite des deux premières égalité pour s'en rendre compte. Démontrons maintenant le point 2 de la proposition. Définissons $P_{10} = q_{C_1^{-1}}(x - \mu_1) + q_{C_0^{-1}}(x - \mu_0)$. On a :

$$\begin{aligned}
P_{10} &= q_{C_1^{-1}}(x-\mu_1) + q_{C_1^{-1}}(x-\mu_0) - q_{C_1^{-1}}(x-\mu_0) + q_{C_0^{-1}}(x-\mu_0) \\
&= 2q_{C_1^{-1}}(x - s_{10}) + \frac{1}{2}q_{C_1^{-1}}(m_{10}) - q_{A_{10}}(x-\mu_0) \\
&\quad \text{(Identité du parallélogramme (1.12) et définition de } S_{10} \text{ et } A_{10}) \\
&= 2q_{C_0^{-1}}(x - s_{10}) + \frac{1}{2}q_{C_0^{-1}}(m_{10}) + q_{A_{10}}(x-\mu_1) \\
&\quad \text{(version symétrique de l'égalité précédente et } A_{10} = -A_{01}) \\
&= 2q_{S_{10}}(x - s_{10}) + \frac{1}{2}q_{S_{10}}(m_{10}) - 2\langle A_{10}m_{10}, x - s_{10}\rangle_{\mathbb{R}^p},
\end{aligned}$$

où la dernière égalité résulte de la moyenne des deux qui la précédent et de l'identité de polarisation (1.13) (même principe que pour obtenir (1.17)). Notons par ailleurs que

$$\mathcal{L}_{10}^Q + \frac{P_{10}}{2} = q_{C_1^{-1}}(x-\mu_1) + \frac{1}{2}\log\left(|det(C_0^{-1}C_1)|\right),$$

et

$$-\mathcal{L}_{10}^Q + \frac{P_{10}}{2} = q_{C_0^{-1}}(x-\mu_0) + \frac{1}{2}\log\left(|det(C_1^{-1}C_0)|\right).$$

Ainsi, puisque

$$\forall i = 1, 0 \quad \log(\frac{dP_i}{dx}) = -\frac{1}{2}\left(q_{C_i^{-1}}(x-\mu_i) + \log(|\det(C_i)|)\right),$$

et

$$-2\log(\frac{dP_{1/2}}{dx}) = q_{S_{10}}(x - s_{10}) + \log(|\det(S_{10}^{-1})|),$$

on obtient

$$f_{10}(x) = -\frac{1}{4}\log\left(|det(S_{10}^2 C_0 C_1)|\right) - \frac{1}{8}q_{S_{10}}(m_{10}) + \frac{1}{2}\langle A_{10}m_{10}, x - s_{10}\rangle_{\mathbb{R}^p}, \tag{1.23}$$

et en substituant $X = S_{10}^{-1/2}U + s_{10}$ à x, on a bien l'inégalité (1.21). L'équation (1.22) résulte elle de (1.15) (proposition précédente).

Démontrons maintenant le point 3 de la proposition. Dans le cas où $C_1 = C_0 = C$, $A_{10} = 0$ et $S_{10} = C^{-1}$. Ainsi, dans ce cas, en notant $\sigma = \frac{1}{2}\|C^{-1/2}m_{10}\|_{\mathbb{R}^p}$ et ξ est une variable aléatoire

réelle gaussienne centrée réduite, on a :

$$\begin{aligned} d_1(P_1,P_0) &= 2e^{-\frac{\sigma^2}{2}}\mathbb{E}\left[\left|\sinh\left(\langle C^{-1/2}m_{10},U\rangle_{\mathbb{R}^p}\right)\right|\right] \\ &= 2e^{-\frac{\sigma^2}{2}}\mathbb{E}\left[|\sinh\sigma\xi|\right] \\ &= e^{-\frac{\sigma^2}{2}}\left(\int_0^\infty \frac{e^{x-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}dx - \int_0^\infty \frac{e^{-x-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}dx\right) \\ &= e^{-\frac{\sigma^2}{2}}\left(e^{\frac{\sigma^2}{2}}P(\xi>-\sigma) - e^{\frac{\sigma^2}{2}}P(\xi>\sigma)\right). \end{aligned}$$

□

Remarque 1.2. *Notons que lorsque $C = C_1 = C_0$, la mesure $P_{1/2} = \gamma_{C,s_{10}}$ joue un rôle symétrique important : elle permet d'obtenir le point 3. Dans le cas où $C_1 \neq C_0$ c'est la mesure $P_{1/2} = \gamma_{S_{10}^{-1},s_{10}}$ qui joue un rôle symétrique important puisque si $X \rightsquigarrow P_{1/2}$ alors d'une part $f_{10}(P_{1/2},X)$ a la même loi que $f_{01}(P_{1/2},X)$ et d'autre part $\mathcal{L}_{10}(X)$ a la même loi que $-\mathcal{L}_{01}(X)$.*

1.5 Comment choisir le seuil d'un test pour lui assurer un niveau α

Supposons que notre test soit de la forme $1_{X>\alpha}$ où X est une variable aléatoire réelle. Si l'on veut connaître le plus grand seuil t_α tel que $P_0(X > t_\alpha) \leq \alpha$ ou tout du moins une valeur inférieure assez proche de la vraie valeur, il faut connaître les quantiles de la distribution de X. Quand on ne les connaît pas, si la distribution de X est trop complexe, on peut chercher à construire notre test différemment pour rapprocher la loi de X (en un certain sens) par une loi dont les quantiles sont connus. L'exemple le plus couramment utilisé est l'approximation normale. Si notre statistique est $X = \sum_{i=1}^n Y_i$ (où les Y_i sont n variables aléatoires identiquement distribuées) et que n est grand, il suffit de centrer et réduire X, de choisir le seuil $z_{1-\alpha}$ (quantile de la loi normale centrée réduite) et de vérifier que l'approximation normale est valable afin d'assurer un seuil qui asymptotiquement vérifie la propriété désirée. Autrement dit le niveau du test ainsi construit tend vers α quand n tend vers l'infini mais il tend peut être vers α par valeurs supérieures et il est toujours préférable de s'assurer par des expérimentations pratiques que le niveau du test ainsi construit n'est pas trop éloigné du niveau recherché.

1.6 Erreur de test, distances et affinités entre mesures

1.6.1 Affinité de test

L'affinité de test est définie pour deux mesures (pas forcement des mesures de probabilité) ν_1 et ν_0 par :

$$A_1(\nu_1,\nu_0) = \int \min(d\nu_1, d\nu_0). \tag{1.24}$$

Lorsque ν_1 et ν_0 sont deux mesures de probabilité, cette dernière quantité est nulle si les deux distributions sont orthogonales et égale à 1 si elles sont égales entre elles ν_1-presque sûrement.

D'une manière générale, on peut dire que l'affinité de test mesure l'écart à l'orthogonalité. Le principal intérêt de cette quantité est contenu dans le théorème suivant.

Theoreme 1.2. *La quantité $A_1(\nu_1, \nu_0)$ est liée à la distance L_1 par*

$$2A_1(\nu_1, \nu_0) = \int (\nu_1 + \nu_0) - |\nu_1 - \nu_0|_1, \tag{1.25}$$

où $|\nu_1 - \nu_0|_1 = \int |d\nu_1 - d\nu_0|$ est la distance L_1 ; et aux erreurs de tests par

$$g_1(\alpha_0, P_1, P_0) = \sup_{t\in[0,\infty]} \left(A_1(P_1, tP_0) - t\alpha_0\right), \tag{1.26}$$

g_1 étant définie par (1.5), et

$$g_2(t, P_1, P_0) = A_1(tP_0, (1-t)P_1), \tag{1.27}$$

g_2 étant définie par (1.6).

Remarque 1.3. *L'affinité de test est donc une quantité duale de la distance L_1 qui permet de mesurer exactement l'erreur associée au meilleur test entre deux hypothèses données. Elle est donc centrale en théorie des tests. Malheureusement son calcul n'est quasiment jamais faisable.*

Notons que le choix particulier de $t = 1$ parmi les minorant

$$A_1(P_1, tP_0) - t\alpha_0 \quad t \in [0, \infty]$$

de (1.26) permet avec (1.25), d'obtenir :

$$g_1(\alpha_0, P_1, P_0) \geq A_1(P_1, P_0) - \alpha_0 = 1 - \alpha_0 - \frac{1}{2}|P_1 - P_0|_1.$$

Démonstration. Pour la démonstration de la première égalité il suffit de faire la différence des deux égalités

$$\int \min(d\nu_1, d\nu_0) + \int \max(d\nu_1, d\nu_0) = \int d\nu_1 + \int d\nu_0, \tag{1.28}$$

et

$$\int \max(d\nu_1, d\nu_0) - \int \min(d\nu_1, d\nu_0) = \int |d\nu_0 - d\nu_1|. \tag{1.29}$$

Pour la deuxième, notons que

$$A_1(\nu_1, \nu_0) = \inf_{\psi\in\Psi} \int \psi d\nu_0 + (1-\psi)d\nu_1,$$

(ce qui résulte du fait que l'infimum en question est atteint pour $\psi = 1_{d\nu_0 \leq d\nu_1}$). On a alors d'une part en posant $\nu_1 = P_1$ et $\nu_0 = tP_0$ et en utilisant (1.8) :

$$g_1(\alpha_0, P_1, P_0) = \sup_{t\in[0,\infty]} \left(A_1(tP_0, P_1) - t\alpha_0\right)$$

et d'autre part en posant $\nu_1 = (1-t)P_1$ et $\nu_0 = tP_0$

$$g_2(t, P_1, P_0) = \inf_{\psi} \int (t\psi dP_0 + (1-t)(1-\psi)dP_1) = A_1(tP_0, (1-t)P_1).$$

□

1.6.2 Distance L_1 et autres distances

Puisque la distance L_1 et l'affinité de test entre deux mesures sont centrales en statistique mais qu'elle sont rarement calculable, il est judicieux d'utiliser d'autres distances ou pseudo distances. Nous introduisons les deux d'entre elles que nous utiliserons et nous donnons les relations entre celles-ci dont nous ferons usage dans la suite du mémoire. Ces distances sont définies ici entre deux mesures de probabilités P_1 et P_0.

La distance de Hellinger. Elle est définie par :

$$h^2(P_1, P_0) = \int (\sqrt{dP_1} - \sqrt{dP_0})^2 = 2(1 - A_2(P_1, P_0))$$

où

$$A_2(P_1, P_0) = \int \sqrt{dP_1 dP_0}$$

est l'affinité de Hellinger. Si t_0 et t_1 sont deux réels positifs, on a clairement

$$A_1(t_1 P_1, t_0 P_0) \leq \sqrt{t_1 t_0} A_2(P_1, P_0) \tag{1.30}$$

(car $\min(a, b) \leq a^{1/2} b^{1/2}$). D'autre part, en remarquant que

$$A_2^2(P_1, P_0) \leq \int \max(P_1, P_0) \int \min(P_1, P_0),$$

et en utilisant (1.28) on a :

$$A_2^2(P_1, P_0) \leq A_1(P_1, P_0)(2 - A_1(P_1, P_0)). \tag{1.31}$$

Des deux comparaisons établies entre $A_1(P_1, P_0)$ et $A_2(P_1, P_0)$ on déduit la proposition suivante.

Proposition 1.4. *[LeCam] Avec les notations précédentes, on a*

$$h^2(P_1, P_0) \leq |P_1 - P_0|_1 \leq 2h(P_1, P_0)\sqrt{1 - \frac{h^2(P_1, P_0)}{4}}. \tag{1.32}$$

Démonstration. La borne inférieure découle de (1.30) avec $t_0 = t_1 = 1$. Par ailleurs, l'équation (1.31) implique

$$\left(1 - \frac{h^2(P_1, P_0)}{2}\right)^2 \leq \left(1 - \frac{d_1(P_0, P_1)}{2}\right)\left(1 + \frac{d_1(P_0, P_1)}{2}\right) = 1 - \frac{d_1^2(P_0, P_1)}{4},$$

dont on déduit la borne supérieure. □

Cette inégalité est due à LeCam. Un des principaux intérêt de la distance h est son lien avec la distance L_1 et le fait suivant. Si pour $j = 1, 0$ $P_j = \otimes_{i=1}^n P_{ji}$ est une mesure produit sur $\mathcal{X}^n$ alors on vérifie facilement que $A_2(P_1, P_0) = \prod_{i=1}^n A_2(P_{1i}, P_{0i})$. La proposition suivante nous sera particulièrement utile dans le Chapitre 2 de la Partie II de ce mémoire.

Proposition 1.5.

$$P_0\left(\frac{dP_1}{dP_0}(X) \geq e^{2x}\right) \leq e^{-x-\frac{h^2(P_1,P_0)}{2}} \tag{1.33}$$

Démonstration. L'inégalité de Markov, et le fait que $\mathbb{E}_0[(\frac{dP_1}{dP_0}(X))^{1/2}] = A_2(P_1, P_0)$ impliquent

$$P_0\left(\left(\frac{dP_1}{dP_0}(X)\right)^{1/2} \geq e^{x}\right) = e^{-x}A_2(P_1, P_0).$$

Le fait que $\log(A_2(P_1, P_0)) \leq -\frac{h^2(P_1,P_0)}{2}$ (qui résulte de l'inégalité de convexité $\log(1-x) \leq -x$) nous permettent alors de conclure. □

La divergence de Kullback. Elle est définie par

$$K(P_1, P_0) = \begin{cases} \int \log\left(\frac{dP_1}{dP_0}\right) dP_1 & si P_1 \ll P_0 \\ \infty & sinon \end{cases}. \tag{1.34}$$

Ce n'est pas une distance, elle n'est pas symétrique.

La divergence du χ^2. Elle est définie par

$$\chi^2(P_1, P_0) = \begin{cases} \int \left(\frac{dP_1}{dP_0} - 1\right)^2 dP_0 & \text{si } P_1 \ll P_0 \\ \infty & \text{sinon} \end{cases}.$$

Ce n'est pas une distance, elle n'est pas symétrique. On remarque que dans le cas où cette divergence est finie, elle vaut

$$\chi^2(P_1, P_0) = \mathbb{E}_0(L^2) - 1,$$

où $\mathbb{E}_0$ est l'espérance sous la loi P_0 et L est le rapport de vraisemblance entre P_1 et P_0. Si pour $j = 1, 0$ $P_j = \otimes_{i=1}^n P_{ji}$ est une mesure produit sur $\mathcal{X}^n$ alors on vérifie facilement que $\mathbb{E}_0(L^2) = \prod_{i=1}^n \mathbb{E}_{0i}(L_i^2)$ (où $\mathbb{E}_{0i}$ est l'espérance sous la loi P_{0i} et L_i est le rapport de vraisemblance entre P_{1i} et P_{0i}). De plus, on montre (cf [75]) la proposition suivante.

Proposition 1.6. *La distance L_1, la distance de Hellinger, la divergence de kullbach et du χ^2 vérifient :*

$$\frac{1}{2}|P_1 - P_0|_1 \leq h(P_1, P_0) \leq \sqrt{K(P_1, P_0)} \leq \sqrt{\chi^2(P_1, P_0)}. \tag{1.35}$$

Chapitre 2

Approche minimax

All we know about the world teaches us that the effects of A and B are always different -in some decimale place- for any A and B. Thus asking "are the effects different ?" is foolish.

Tukey

Dans ce chapitre, nous introduisons la problématique des tests minimax. Nous détaillons l'étude d'une alternative composite finie en donnant une preuve du lemme minimax ainsi qu'une interprétation géométrique de cette preuve. Nous montrons comment tenir compte des symétries dans les problèmes de tests. A partir d'exemples simples, nous introduisons la notion de séparation d'hypothèses. Nous illustrons par des exemples simples mais fondamentaux le calcul de ces frontières de séparation. Nous expliquons en quoi la taille de l'alternative appelle une réduction de dimension.

2.1 Généralités

Le plus souvent, $\mathcal{P}_0$ et $\mathcal{P}_1$ ne sont pas réduits à un élément. Il faut alors, pour parvenir à utiliser ce qui a été fait au chapitre précédent, se donner les moyens d'étudier de manière globale (uniformément sur $\mathcal{P}_0 \cup \mathcal{P}_1$) les performances d'un test ψ. Autrement dit, il s'agit de définir des quantités du type $\alpha(\psi, \mathcal{P}_0)$ et $\beta(\psi, \mathcal{P}_1)$ et de les étudier. Il y a deux méthodes, utilisant des quantités locales introduites dans le cadre des tests simples, pour définir ces mesures d'erreur globales.

1. La première correspond à introduire deux mesures de probabilité π_0 et π_1 respectivement sur $\mathcal{P}_0$ et $\mathcal{P}_1$, les quantités globales mesurant l'erreur sont alors :

$$\alpha(\psi, \mathcal{P}_0) = \mathbb{E}_{\pi_0}[\alpha(\psi, P_0)] \quad \text{et} \quad \beta(\psi, \mathcal{P}_1) = \mathbb{E}_{\pi_1}[\beta(\psi, P_1)].$$

   Cette démarche est dite Bayesienne (puisque trouver un estimateur qui minimise l'erreur globale ainsi introduite correspond à faire de l'inférence en applicant le principe de Bayes),

et les mesures π_0 et π_1 sont dites a priori.

2. La seconde démarche, dite minimax, consiste à prendre la pire des erreurs comme représentante de l'erreur globale :

$$\alpha(\psi, \mathcal{P}_0) = \sup_{P_0 \in \mathcal{P}_0} \alpha(\psi, P_0) \quad \text{et} \quad \beta(\psi, \mathcal{P}_1) = \sup_{P_1 \in \mathcal{P}_1} \beta(\psi, P_1).$$

Dans toute la suite, nous noterons

$$\Psi_{\alpha_0} = \{\psi \in \Psi : \alpha(\psi, \mathcal{P}_0) \leq \alpha_0\}.$$

Dans le cas de la démarche bayesienne, on peut définir les mesures de probabilité :

$$\forall A \in \mathcal{A} \;\; P_{\pi_i}(A) = \mathbb{E}_{\pi_i}(P(A)) \;\; i = 0, 1 \; .$$

On obtient

$$\beta^{\pi_1}(\psi, \mathcal{P}_1) = \beta(\psi, P_{\pi_1}) \quad \text{et} \quad \alpha^{\pi_0}(\psi, \mathcal{P}_0) = \alpha(\psi, P_{\pi_0}).$$

D'après le Lemme de Neymann Pearson, pour tout $\alpha_0 \in [0, 1]$, il existe $t_0 \in [0, \infty]$ tel que le meilleur test $\psi^*(t_0, \pi_0, \pi_1)$ soit donnée par :

$$\psi^*(t_0) = \begin{cases} 1 & \text{si } dP^\pi > t_0 dP_0 \\ c(t_0) & \text{si } dP^\pi = t_0 dP_0 \\ 0 & \text{si } dP^\pi < t_0 dP_0 \end{cases} .$$

2.2 Hypothèse nulle simple contre alternative composite finie

Problématique et notations. Nous allons détailler l'analyse théorique du cas particulier d'hypothèses du type :

$$\mathcal{P}_0 = \{P_0\} \quad \mathcal{P}_1 = \{P_1, \ldots, P_N\}.$$

D'une part, ce cas permettra de saisir le lien entre le risque minimax et le risque bayesien lorsque la structure des hypothèses est assez simple pour se prêter à une description élémentaire et géométrique. D'autre part nous pourrons illustrer par des exemples l'influence de N (correspondant à la taille de l'ensemble des hypothèses alternatives) sur la difficulté du problème.
Notons tout de suite que $\mathcal{P}_1$ est dominé par une mesure finie (car fini). Afin de mettre en valeur l'aspect géométrique de la démarche, nous adopterons les notations suivantes :

$$\beta_i(\psi) = \beta(\psi, P_i) \quad (\beta_i(\psi))_{i=1\ldots,N} = B(\psi) \in \mathbb{R}^N.$$

D'après le Théorème 1.1, pour tout $i \in \{1, \ldots, N\}$ le sous-ensemble de $[0, 1]$ $\beta_i(\Psi_{\alpha_0})$ est un convexe fermé. On peut donc affirmer que l'ensemble

$$\mathcal{R}_N(\alpha_0) = \left\{B(\psi) \in [0, 1]^N \quad \psi \in \Psi_{\alpha_0}\right\},$$

est un domaine convexe fermé, ce domaine est l'ensemble des erreurs de seconde espèce associées à tous les tests ψ ayant une erreur de première espèce inférieure ou égale à α_0. Le point $U = (1, \ldots, 1)$ est dans $\mathcal{R}_N(\alpha_0)$.

Puisque $\mathcal{P}_1$ est finie, une loi a priori π sur $\mathcal{P}_1$ est une loi discrète finie : $(\pi_1, \ldots, \pi_N) \in \Pi_N = \{x \in [0,1]^N \ : \ \sum x_i = 1\}$, autrement dit les erreurs de seconde espèce, sont selon la démarche :

$$\beta^\pi(\psi, \mathcal{P}_1) = \sum_{i=1}^N \pi_i \beta(\psi, P_i) = \langle \pi, B \rangle_{\mathbb{R}^N} \text{ (cadre bayesien) ,}$$

$$\text{ou } \beta(\psi, \mathcal{P}_1) = \max_{i=1,\ldots,N} \beta(\psi, P_i) = \max_{i=1,\ldots,N} \beta_i(\psi) \text{ (cadre minimax) .}$$

Lemme Minimax. Avant d'énoncer le lemme minimax, nous rappelons la définition d'invariance d'un problème de test.

Définition 2.1 (Invariance)**.** *Soit $f : \mathcal{P}_1 \to \mathcal{P}_1$ une transformation donnée. Un problème de test des hypothèses $\mathcal{P}_0 = \{P_0\}$ contre $\mathcal{P}_1 = \{P_1, \ldots, P_N\}$ est dit* ***invariant par l'action de*** *f si*

$$\{(\beta(\psi, P_i))_{i\in\{1,\ldots,N\}} \ \ \psi \in \Psi_{\alpha_0}\} = \{(\beta(\psi, f(P_i)))_{i\in\{1,\ldots,N\}} \ \ \psi \in \Psi_{\alpha_0}\}.$$

Le lemme suivant est au coeur de la solution au problème minimax, il établit le lien entre risque minimax et risque bayesien. Notons que ce lemme donné dans un cadre un peu plus général (la démonstration ne diffère pas) est la base du théorème minimax obtenu par Ky Fan en 1953 [33]. Le théorème minimax de Ky Fan se déduit du lemme suivant par un argument topologique.

Lemme 2.1 (Geometrie et minimax-bayesien)**.** *1. Le risque minimax est obtenu pour l'a priori π^* qui rend maximal le risque bayesien, cet a priori est appelé « a priori le moins favorable » . Autrement dit*

$$\inf_{\psi\in\Psi_{\alpha_0}} \beta^{\pi^*}(\psi, \mathcal{P}_1) = \sup_{\pi\in\Pi} \inf_{\psi\in\Psi_{\alpha_0}} \beta^\pi(\psi, \mathcal{P}_1) = \inf_{\psi\in\Psi_{\alpha_0}} \max_{i=1,\ldots,N} \beta_i(\psi) = \inf_{\psi\in\Psi_{\alpha_0}} \beta(\psi, \mathcal{P}_1).$$

2. Si le problème est invariant par toute permutation des hypothèses alternatives alors l'a priori $\pi^* = (1/N, \cdots, 1/N)$ réalise le risque minimax :

$$\inf_{\psi\in\Psi_{\alpha_0}} \beta(\psi, \mathcal{P}_1) = \inf_{\psi\in\Psi_{\alpha_0}} \beta^{\pi^*}(\psi, \mathcal{P}_1).$$

Démonstration. Si l'on définit $S(x_1, \ldots, x_N) = \max_i(x_i)$ alors le risque minimax est simplement le minimum de la fonction sous-linéaire[1] S sur $\mathcal{R}_N(\alpha_0)$:

$$\inf_{\psi\in\Psi_0} \max_{i=1,\ldots,N} \beta(\psi, P_i) = \inf_{X=(x_1,\ldots,x_N)\in\mathcal{R}_N(\alpha_0)} S(X). \tag{2.1}$$

Nous allons utiliser le théorème de Mazur-Orliz Nous l'énonçons ici et renvoyons à [67] pour un énoncé et une preuve de ce théorème.

Theoreme 2.1. *Soit $S : E \to R$ une application sous linéaire, $C \subset E$ un sous-ensemble convexe non vide. Alors, il existe une forme linéaire L sur E telle que*

$$\forall x \in E \ \ S(x) \geq L(x) \ \ \textit{et} \ \inf_{y\in C} L(y) = \inf_{y\in C} S(y).$$

[1] Une application S est sous-linéaire si pour tout x, y $S(x+y) \leq S(x) + S(y)$ et pour tout scalaire $\lambda > 0$ $S(\lambda x) = \lambda S(x)$

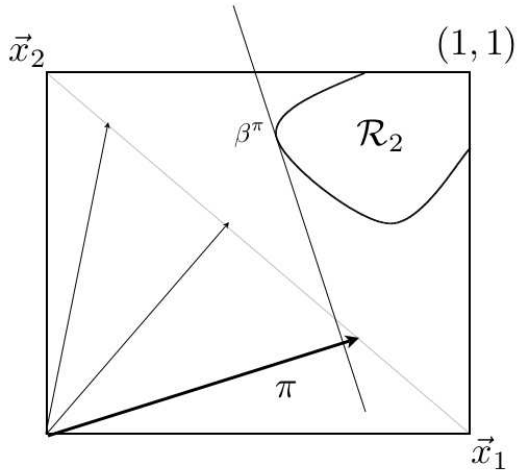


FIG. 2.1 – Risque bayesien β^π associé à l'a priori π quand $N = 2$. A un test ψ de niveau inférieur ou égal à α_0 correspond un point de $\mathcal{R}_2$, qui a pour coordonnées $(\beta_1(\psi), \beta_2(\psi))$.

Ici, E est tout simplement $\mathbb{R}^N$ et $C = \mathcal{R}_N(\alpha_0)$, ainsi, il existe une forme linéaire L telle que

$$S(x) \geq L(x) \ \ \forall x \in \mathbb{R}^N \quad \text{et} \quad \inf_{X=(x_1,\ldots,x_N)\in\mathcal{R}_N(\alpha_0)} S(X) = \inf_{X=(x_1,\ldots,x_N)\in\mathcal{R}_N(\alpha_0)} L(X). \tag{2.2}$$

Notons $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^N$. Puisque S est la fonction maximum, (2.2) impose que $S(\mathbf{1}) = 1 \geq L(\mathbf{1})$ et $S(-\mathbf{1}) = -1 \geq L(-\mathbf{1}) = -L(\mathbf{1})$. Ainsi L est une combinaison convexe (c'est-à-dire $L(\mathbf{1}) = 1$). En d'autres termes, il existe $\pi^* \in \Pi_N$ tel que $L(x) = \langle \pi^*, x \rangle_{\mathbb{R}^N}$. En combinant (2.1) et (2.2) on peut finalement obtenir

$$\inf_{\psi\in\Psi_0} \max_{i=1,\ldots,N} \beta(\psi, P_i) = \inf_{X=(x_1,\ldots,x_N)\in\mathcal{R}_N(\alpha_0)} \langle \pi^*, X \rangle_{\mathbb{R}^N} = \inf_{\psi\in\Psi_0} \beta^{\pi^*}(\psi). \tag{2.3}$$

Pour conclure, notons que l'inégalité

$$\sup_{\pi\in\Pi} \inf_{\psi\in\Psi_{\alpha_0}} \beta^\pi(\psi, \mathcal{P}_1) \leq \inf_{\psi\in\Psi_{\alpha_0}} \max_{i=1,\ldots,N} \beta_i(\psi)$$

est triviale. Aussi, puisque d'après l'équation (2.3) l'égalité est réalisée pour $\pi = \pi^*$, la démonstration de la première partie de la proposition est achevée.

Pour la deuxième partie de la proposition il faut utiliser le fait que (2.2) nous indique en quoi les propriétés de $\mathcal{R}_N(\alpha_0)$ impliquent les propriétés de L (et donc de π^*). Si par exemple le problème de test est invariant par permutation des hypothèses i et j cela signifie que $\mathcal{R}_N(\alpha_0)$ est invariant par l'action de σ_{ij} l'application permutant les coordonnées i et j d'un vecteur de $\mathbb{R}^n$, et donc

$$\inf_{X\in\mathcal{R}_N(\alpha_0)} \langle \pi^*, X \rangle_{\mathbb{R}^N} = \inf_{X\in\mathcal{R}_N(\alpha_0)} \langle \pi^*, \sigma_{ij}(X) \rangle_{\mathbb{R}^N} = \inf_{X\in\mathcal{R}_N(\alpha_0)} \langle \sigma_{ij}(\pi^*), X \rangle_{\mathbb{R}^N}$$

$$\text{et } \forall X \in \mathbb{R}^N, \ S(X) \geq \langle \sigma_{ij}(\pi^*), X \rangle_{\mathbb{R}^N}$$

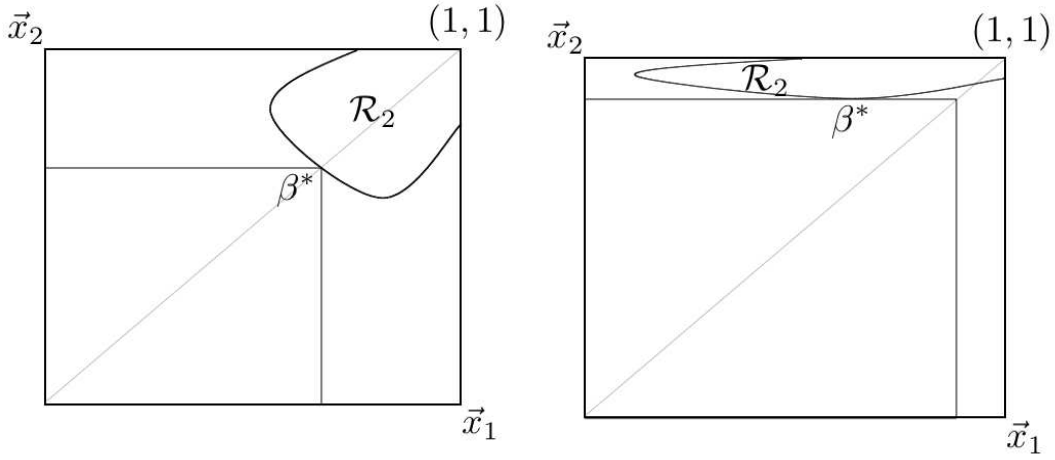


FIG. 2.2 – Deux cas de figure quant à la position de $\mathcal{R}_2$. A un test ψ de niveau inférieur ou égal à α_0 correspond un point de $\mathcal{R}_2$, qui a pour coordonnées $(\beta_1(\psi), \beta_2(\psi))$

Ainsi $\sigma_{ij}(\pi^*)$ est un a priori le moins favorable et toute combinaison convexe de $\sigma_{ij}(\pi^*)$ et π^* l'est aussi. En particulier il existe un a priori le moins favorable $\tilde{\pi}^*$ tel que $\tilde{\pi}^*_i = \tilde{\pi}^*_j$. La deuxième partie de la proposition résulte alors du fait que $\sum_{i=1}^N \tilde{\pi}^*_i = 1$ et que le problème de test est supposé être invariant par n'importe quelle permutation. □

Interpretation géométrique. Nous allons maintenant donner une description géométrique de l'erreur bayesienne d'une part et minimax d'autre part.

Soit H_π l'hyperplan orthogonal à un vecteur $\pi \in \Pi_N$, qui contient des points de $\mathcal{R}_N$ et dont la distance à 0 est minimale[2]. Le risque minimum bayesien correspondant est la distance de l'hyperplan H_π à 0 (voir Figure 2.1) :

$$\inf_{\psi \in \Psi_0} \beta^\pi(\psi, P) = \inf_{B \in \mathcal{R}_N} \langle \pi, B \rangle_N = d(H_\pi, 0). \tag{2.4}$$

La procédure minimax a elle aussi une interprétation géométrique simple : le risque minimax est égal à la longueur r^* du coté du plus petit hypercube $c(r^*) \subseteq \mathbb{R}^N_+$ de sommet 0 qui a une intersection non vide avec $\mathcal{R}_N$. (Les lignes de niveaux de S sont des hypercubes, voir Figure 2.2).

Valeur du risque minimax. On remarque que

$$\beta^\pi(\psi, \mathcal{P}_1) = 1 - \mathbb{E}_{\pi^*}[\mathbb{E}_{P_i}[\psi]] = 1 - \mathbb{E}_Q[\psi] \quad Q = \sum_{i=1}^N \pi^*_i P_i,$$

[2]Puisque $\mathcal{R}_N(\alpha_0)$ est un convexe fermé, l'hyperplan H_π existe et est unique (la véritable de raison de l'existence relève du théorème de Mazur-Orlicz)

et il est donc possible de se ramener au cas du test des hypothèses simples $\{P_0\}$ et $\{Q\}$. Ainsi, il existe $t_0 > 0$ tel que

$$\text{Arginf}_{\psi\in\Psi_{\alpha_0}} \max_{i=1,\dots,N} \beta(\psi, P_i) = \psi^*(t_0) = \begin{cases} 1 & \text{si } dQ > t_0 dP_0 \\ c(t_0) & \text{si } dQ = t_0 dP_0 \\ 0 & \text{si } dQ < t_0 dP_0 \end{cases}.$$

On peut finalement résumer la série de résultats obtenus par les égalités suivantes :

$$\inf_{\psi\in\Psi_0} \max_{i=1,\dots,N} \beta^\pi(\psi) = \sup_{\pi\in\Pi} \inf_{\psi\in\Psi_{\alpha_0}} \beta^\pi(\psi) = \sup_{P\in[\mathcal{P}_1]} \inf_{\psi\in\Psi_{\alpha_0}} \beta(\psi, P), \text{ et} \tag{2.5}$$

$$\inf_{\psi\in\Psi_0} \max_{i=1,\dots,N} \beta_i(\psi) = \sup_{P\in[\mathcal{P}_1]} \sup_{t\in[0,\infty]} \{A_1(P, tP_0) - t\alpha_0\} = \sup_{t\in[0,\infty]} \{A_1(Q, tP_0) - t\alpha_0\}. \tag{2.6}$$

Le lemme suivant est une conséquence de ce résultat dont nous ferons usage par la suite.

Lemme 2.2.

$$\ln\left(\chi^2(Q, P_0) + 1\right) \geq \mathcal{C}\left(\alpha_0 + \inf_{\psi\in\Psi_0} \max_{i=1,\dots,N} \beta_i(\psi)\right), \tag{2.7}$$

où $\mathcal{C}$ est la fonction décroissante sur $[0, 1]$ définie par

$$\mathcal{C}(x) = \ln(1 + 4(1 - x)^2). \tag{2.8}$$

Démonstration. Il suffit d'utiliser (1.25) (égalité liant la distance L_1 et A_1), et (1.6) (inégalité liant la distance L_1 et la divergence du χ^2) en fixant $t = 1$. □

Dans le cas où l'on cherche à contrôler $t\alpha(\psi) + (1 - t)\beta(\psi)$, des calculs similaires peuvent être effectués, et on a alors

$$\inf_{\psi\in\Psi} \max_{i=1,\dots,N} (t\alpha(\psi) + (1 - t)\beta(\psi)) = \sup_{P\in[\mathcal{P}_1]} A_1((1 - t)P_1, tP_0) = A_1((1 - t)Q, tP_0).$$

2.3 Application : Emergence de la dimension, nécessité de la séparation

Nous allons traiter ici deux exemples fondamentaux illustrant la vitesse à laquelle le nombre d'hypothèses N de l'alternative peut provoquer un "rapprochement" de l'alternative et de l'hypothèse nulle.

2.3.1 Premier exemple

Le premier exemple est emprunté à Burnashev et al. ([18]). Rappelons que si C est une matrice symétrique définie positive sur $\mathbb{R}^N$ et $\mu \in \mathbb{R}^N$ alors $\gamma_{C,\mu}$ est la mesure gaussienne de moyenne μ et de covariance C.

Supposons que l'on observe un vecteur $f \in \mathbb{R}^N$ dans un bruit gaussien $\xi \rightsquigarrow \gamma_N$ (γ_N est la mesure gaussienne centrée réduite sur $\mathbb{R}^N$) :

$$X_i = f_i + \xi_i, \quad i = 1, \dots, N$$

On veut tester les hypothèses :

$$H_0 \ : \ f = 0 \quad \text{contre } f \in \rho V(1),$$

où ρ est un réel positif, $V(1) = \{e^i, \ i = 1 \ldots, N\}$ (ensemble des points extrémaux du simplexe en dimension N, $e^i = (0, \ldots, 0, 1, 0, \ldots, 0)$ où le « 1 » est à la i^{eme} position), et $\rho V(1) = \{\rho f \ : f \in V(1)\}$.

Nous sommes donc dans le cadre du lemme minimax avec $P_i = \gamma_{I_N, \rho e^i}$. L'invariance par changement des hypothèses impose que $\pi^* = (1/N, \ldots, 1/N)$ et donc

$$\frac{dQ}{dP_0}(X) = \frac{1}{N} \sum_{i=1}^{N} \frac{dP_i}{dP_0}(X) = \frac{1}{N} \sum_{i=1}^{N} e^{-\frac{\rho^2}{2} + \rho X_i}. \tag{2.9}$$

Le test minimax est alors (si l'on cherche à contrôler l'erreur de première espèce) :

$$\psi^*(t_0) = \begin{cases} 1 & \text{si } \frac{dQ}{dP_0}(X) > t_0 \\ c(t_0) & \text{si } \frac{dQ}{dP_0}(X) = t_0 \\ 0 & \text{si } \frac{dQ}{dP_0}(X) < t_0 \end{cases}, \tag{2.10}$$

t_0 étant le plus grand réél permettant de maintenir le test au niveau α_0, et on a la proposition qui suit.

Proposition 2.1 (Burnashev et Begmatov [18])**.** *Supposons que la dimension du problème est liée à ρ (la distance de l'alternative à l'origine) par :*

$$N(\rho, h, r) = \exp\{h\rho + r\rho^2\},$$

et que $\alpha_0 < 1$. On a alors :

$$\inf_{\psi \in \Psi_{\alpha_0}} \max_{i=1,\ldots,N} \beta(\psi, P_i) \to \begin{cases} 0 & si \ r < 1/2 \\ 1 & si \ r > 1/2 \end{cases} \quad quand \ \rho \to \infty.$$

Pour la démonstration, voir Burnashev et Begmatov [18].

Pour que le test minimax soit efficace, il faudra que les moyennes des hypothèses alternatives soient à une distance au moins $\sqrt{(1/2 + \epsilon) \log N}$ de 0 pour un $\epsilon > 0$. Si cela est le cas, le test correspondant sera parfaitement efficace asymptotiquement. Si $-1/2 < \epsilon < 0$ les hypothèses ne seront pas distinguables asymptotiquement. Cette distance frontière est la distance de séparation minimax asymptotique des hypothèses que nous allons définir après avoir expliqué en quoi le test minimax donné par (2.10) est basé sur une réduction de dimension.

2.3.2 Nécessité de la réduction de dimension

La statistique de test donnée par (2.9) cache une procédure de réduction de dimension. En effet, dans la somme

$$\frac{1}{N} \sum_{i=1}^{N} e^{-\frac{\rho^2}{2} + \rho X_i},$$

on donne un poids à chaque observation X_i. Ce poids est donné par $e^{\rho X_i}$. Ainsi, relativement à celles qui ont une grande valeur, les observations ayant une valeur assez petite ne contribuent presque pas à augmenter $\frac{dQ}{dP_0}(X)$. On peut affirmer que les petites observations sont presque seuillées (c'est-à-dire mises à zero).

2.3.3 Séparation minimax d'hypothèses

Nous introduisons une définition de séparation minimax d'hypothèses dans un cadre assez simple. Supposons que f est une suite de carré sommable dont les n premières coordonnées sont observées dans un bruit gaussien i.i.d de loi normale centrée de variance σ^2 :

$$Y_i = f_i + \sigma\epsilon_i \;\; i \in \{1, \dots, n\}. \tag{2.11}$$

On cherche à tester les hypothèses

$$H_0 \;\; : \;\; f = 0 \;\; \text{contre} \;\; H_1 \;\; : \; f \in \mathcal{P}_1(\mathcal{F}, \rho) = \{f \in \mathcal{F} \text{ et } \|f\|_2 \geq \rho\}, \tag{2.12}$$

où $\mathcal{F}$ est une partie de l^2 telle que l'ensemble des hypothèses alternatives soit non vide (mais pas nécessairement fini ou dénombrable).

Définition 2.2. *On appelle vitesse minimax de séparation dans le test des hypothèses définies par (2.12) la quantité*

$$\rho_n(\mathcal{F}, \alpha_0, \beta_0, \sigma) = \inf\left\{\rho > 0 \;\; tq \;\; \inf_{\psi \in \Psi_{\alpha_0}} \beta(\psi, \mathcal{P}_1(\mathcal{F}, \rho)) \leq \beta_0\right\}. \tag{2.13}$$

La vitesse minimax asymptotique est la fonction définie à une constante indépendante de n près, par

$$\rho^a(\mathcal{F}, \alpha_0, \beta_0, \sigma) = \inf\left\{\rho > 0 \;\; tq \;\; \inf_{\psi \in \Psi_{\alpha_0 + o_n(1)}} \beta(\psi, \mathcal{P}_1(\mathcal{F}, \rho)) \leq \beta_0 + o_n(1)\right\}. \tag{2.14}$$

Nous allons donner un deuxième exemple plus général illustrant le lien entre frontière de séparation et dimension (effective) du problème de test.

2.3.4 Deuxième exemple.

Le deuxième exemple est une adaptation d'une petite partie des travaux de Yannick Baraud [7].

Supposons maintenant que l'ensemble des alternatives soit composé des vecteurs qui sont somme de exactement k éléments de la base canonique f_i multipliés par $\pm\rho$. Définissons pour cela

$$\bar{e}^{J_k} = \sum_{i \in J_k} \epsilon_i e^i \;\; \text{et} \;\; \tilde{V}(k) = \left\{\bar{e}^{J_k} \;\; J_k \subset [1, n] \quad \text{Card}(J_k) = k \quad (\epsilon_i)_i \in \{-1, +1\}^n\right\}. \tag{2.15}$$

Notre problème est alors, au vu de

$$Y_i = f_i + \sigma\xi_i,$$

de tester les hypothèses :

$$H_0 \;\; f = 0 \quad \text{contre} \quad f \in \rho\tilde{V}(k).$$

L'invariance par rotation de la mesure gaussienne de covariance identité et de moyenne nulle implique que le problème de test est invariant par toute permutation des hypothèses alternatives. En effet, si $\psi \in \Psi_0$ et R une matrice orthogonale sur $\mathbb{R}^N$ alors $\psi \circ R \in \Psi_0$. L'a priori le moins favorable est donc uniforme. Notons $(\omega_i)_{i=1,\dots,n}$ des variables aléatoires de Rademarcher :

indépendantes et valant -1 et $+1$ avec une probabilité $1/2$. Ainsi, si l'on utilise le fait que pour ces variables et une fonction f de k variables $\frac{1}{2^k}\sum_{\eta\in\{-1,1\}^k} f(\eta_1 x_1,\dots,\eta_k x_k) = E[f(\omega_1 x_1,\dots,\omega_k x_k)]$, on a par symétrie

$$\frac{dQ}{dP_0}(X) = \frac{1}{C_n^k}\sum_{J_k}\mathbb{E}\left[e^{-\frac{\rho^2|J_k|}{2\sigma^2}+\frac{\rho}{\sigma}\sum_{i\in J_k}\omega_i\frac{X_i}{\sigma}}|X\right] \quad = \frac{1}{C_n^k}\sum_{J_k}\prod_{i\in J_k}\mathbb{E}[W_i^{\omega_i}(\rho,\sigma)|X],$$

où

$$W_i^{\omega_i}(\rho,\sigma) = e^{-\frac{\rho^2}{2\sigma^2}+\frac{\rho}{\sigma}\omega_i\frac{X_i}{\sigma}}.$$

On a finalement :

$$\frac{dQ}{dP_0}(X) = \frac{e^{-\frac{k\rho^2}{2\sigma^2}}}{C_n^k}\sum_{J_k}\prod_{i\in J_k}\cosh(\frac{\rho}{\sigma}\frac{X_i}{\sigma}),$$

et le test minimax est :

$$\psi^*(t_0) = \begin{cases} 1 & \text{si } \frac{dQ}{dP_0}(X) > t_0 \\ c(t_0) & \text{si } \frac{dQ}{dP_0}(X) = t_0 \\ 0 & \text{si } \frac{dQ}{dP_0}(X) < t_0 \end{cases}, \tag{2.16}$$

t_0 étant le plus petit réél permettant de maintenir le test au niveau α_0.

Proposition 2.2 (Baraud [7])**.** *Si l'on note* $\mathcal{C}^* = \mathcal{C}\left(\alpha_0 + \inf_{\psi\in\Psi_0}\max_{i=1,\dots,N}\beta^\pi(\psi)\right)$

$$\rho^2 \geq \ln\left(1 + \mathcal{C}^*\frac{n}{k^2} + \sqrt{2\mathcal{C}^*\frac{n}{k^2} + \left(\mathcal{C}^*\frac{n}{k^2}\right)^2}\right)\sigma^2. \tag{2.17}$$

Dans le cas où $k = n$,

$$\rho^2 \geq \sigma^2\sqrt{2\frac{\mathcal{C}^*}{k}}.$$

Remarque 2.1. *Donnons plus exactement la formulation faite par Yannick Baraud. Les points de* $\rho\tilde{V}(k)$ *sont à une distance* $\sqrt{k}\rho$ *de* 0*. Ainsi, si* $\mathcal{F}$ *est l'espace vectoriel engendré par les éléments de* $\tilde{V}_k$*, le résultat précédent nous dit que dans le cas où* $k < n$,

$$\rho_n(\mathcal{F},\alpha_0,\beta_0,\sigma)^2 \geq k\ln\left(1 + \mathcal{C}(\alpha_0+\beta_0)\frac{n}{k^2} + \sqrt{2\mathcal{C}(\alpha_0+\beta_0)\frac{n}{k^2} + \left(\mathcal{C}(\alpha_0+\beta_0)\frac{n}{k^2}\right)^2}\right)\sigma^2 = \rho_{k,n}^2, \tag{2.18}$$

et dans le cas où $k = n$,

$$\rho_n(\mathcal{F},\alpha_0,\beta_0,\sigma)^2 \geq \sqrt{2k\mathcal{C}^*}.$$

Démonstration. Yannick Baraud ([7] en annexe de l'article) obient la majoration suivante

$$\mathbb{E}_0\left[\frac{dQ}{dP_0}^2\right] \leq e^{k\ln(1+\frac{k}{n}(\cosh(\frac{\rho^2}{\sigma^2})-1))},$$

ainsi, on a en utilisant le fait que $\ln(1+x)\leq x$:

$$\frac{n}{k^2}\ln\left(\mathbb{E}_0\left[\frac{dQ}{dP_0}^2\right]\right) + 1 \leq \cosh\left(\frac{\rho^2}{\sigma^2}\right),$$

soit avec (2.7)

$$\frac{n}{k^2}\mathcal{C}^* + 1 \leq \cosh\left(\frac{\rho^2}{\sigma^2}\right).$$

Le fait que sur $[0, \infty[$ la fonction $\cosh(x)$ soit inversible d'inverse $g(y) = \sqrt{\ln(y + \sqrt{y^2 - 1})}$ permet d'obtenir l'inégalité

$$g\left(\max\left(1, \frac{n}{k^2}\mathcal{C}^* + 1\right)\right)\sigma^2 \leq \rho^2$$

dont résulte la proposition. □

2.3.5 Nécessité de la réduction de dimension

Puisque pour un vecteur donné de l'alternative, k composantes seulement sont non nulles, on peut parler pour k de dimension effective du problème. L'espace engendré par les vecteurs de l'alternative est de dimension n. C'est le rapport k/n qui détermine l'intérêt d'une réduction de dimension. Dans le cas limite où $k = 1$, on peut effectuer la même remarque que celle en faveur de la réduction de dimension à la suite de l'exemple 1. Dans le cas où $k = n$ alors

$$\frac{dQ}{dP_0} \propto \prod_{i=1}^{n} \cosh\left(\frac{\rho X_i}{\sigma^2}\right).$$

Ainsi toutes les observations sont considérées comme d'égale importance. On peut donc noter que dans ce cas la procédure minimax ne cache pas de procédure de réduction de dimension. Dans les cas intermédiaire, on a :

$$\frac{dQ}{dP_0} \propto \sum_{J_k} \prod_{i \in J_k} \cosh\left(\frac{\rho X_i}{\sigma^2}\right).$$

Il est difficile de voir directement dans cette expression, quels sont exactement les cas où l'on peut parler de procédure de réduction de dimension. On peut affirmer deux choses. Premièrement, ces cas intermédiaires constituent un continuum entre réduction de dimension et pas de réduction de dimension. Deuxièmement, dans les cas intermédiaires où l'on peut parler de réduction de dimension (typiquement lorsque k est petit), la statistique de test (et donc la réduction de dimension) dépend assez nettement de la connaissance de k. Nous le verrons par la suite, la méconnaissance de k est (en dehors du choix du seuil t_0) à l'origine de certains problèmes importants dans la pratique.

2.3.6 Deux exemples de vitesse de séparation

Nous allons reprendre deux exemples de l'article de Yannick Baraud [7] pour donner deux applications des résultats précédents et une borne inférieure des vitesses minimax sur les boules de Besov. Dans le chapitre suivant nous décrivons des tests qui permettent d'atteindre asymptotiquement ces vitesses.

Il est possible d'utiliser plusieurs type de boules $\mathcal{F}(R)$ de l^2. Le choix d'une boule particulière peut être motivées par une volonté de modélisation. Une boule donnée aura de propriétés d'approximation adaptées à un problème donné. La première annexe de ce mémoire est consacrée à quelques définitions et rappels liés à la théorie de l'approximation et nous n'en parlons donc pas ici. Nous nous contentons d'étudier les deux applications suivantes.

Ellipsoide l^p pour $0 < p < 2$. Soit $a \in \mathbb{R}^{\mathbb{N}}$ décroissant vers 0 et telle que $a_1 = 1$. Nous rappelons que l'ellipsoide l^p de rayon R associée à a est défini par

$$\mathrm{E}_{a,p}(R) = \left\{ f \in l^2,\ \sum_{k>0} \left| \frac{f_k}{a_k} \right|^p \leq R^p \right\}. \tag{2.19}$$

Si x est un réel positif, nous noterons $\lceil x \rceil$ la partie entière de x. Nous noterons, pour $D \neq 0$,

$$r_D^2 = \min\left(\rho_{\lceil \sqrt{D} \rceil, D}, R^2 a_D^2 \lceil \sqrt{D} \rceil^{1-2/p} \right) \text{ et } \rho_{a,p,R}^2 = \sup_{D \in \mathbb{N}^*} r_D^2, \tag{2.20}$$

où $\rho_{k,n}$ est définit par (2.18). On a la proposition suivante

Proposition 2.3 (Baraud [7]). *Si $\mathcal{F} = \mathrm{E}_{a,p}(R)$, alors on a :*

$$\rho_n(\mathcal{F}, \alpha_0, \beta_0, \sigma) \geq \rho_{a,p,R}.$$

La démonstration de cette proposition donnée par Baraud [7] est instructive. Il s'agit de montrer que l'on peut inclure $r_D \tilde{V}_D(\lceil \sqrt{D} \rceil)$ dans $\{f \in \mathcal{E}_{a,p}(R)\ tq\ \|f\|_2 = r_D\}$. En effet, puisque $r_D \leq \rho_{\lceil \sqrt{D} \rceil, D}$ ceci implique que $\inf_{\psi \in \Psi_{\alpha_0}} \beta(\psi, \mathcal{P}_1(\mathcal{E}_{a,p}(R), r_D)) > \beta_0$ et donc $\rho_n(\mathcal{E}_{a,p}(R), \sigma) \geq r_D$. On obtient alors par passage au sup l'inégalité

$$\inf_{\psi \in \Psi_{\alpha_0}} \beta(\psi, \mathcal{P}_1(\mathrm{E}_{a,p}(R), \rho_{a,p,R})) > \beta_0,$$

($\mathcal{P}_1(\mathrm{E}_{a,p}(R), \rho_{a,p,R})$ est définie par (2.12), et $\rho_{a,p,R}$ par (2.20)). On en déduit le résultat voulu.

Ellipsoide de Besov. Nous donnons ici une borne inférieure de la vitesse minimax sur les ellipsoides de Besov $\mathcal{B}_{s',p,p}$ définies par

$$\mathcal{B}_{s',p,q}(R) = \left\{ f \in l^2,\ \sum_{j \geq 0} \left[2^{jsp} \sum_{k=2^j}^{2^{j+1}-1} |f_k|^p \right] \leq R^p \right\}. \tag{2.21}$$

Proposition 2.4 (Baraud [7]). *Soient $p < 2$, $s > 0$ et $s'' = s - 1/4 + 1/(2p)$. Supposons que $\sigma^2 < R^2$ et que $\alpha_0 + \beta_0 \leq 0.29$, alors*

$$\rho_n(\mathcal{B}_{s,p,p}, \alpha_0, \beta_0, \sigma) \geq 2^{-4s''} \min\left(R^{2/(1+4s'')} \sigma^{8s''/(1+4s'')}, \sqrt{n}\sigma^2 \right). \tag{2.22}$$

La preuve donnée dans Baraud [7] repose sur une application de la proposition précédente.

2.3.7 Nécessité (ou non) de réduire la dimension

Remarquons que $\min(R^{2/(1+4s'')} \sigma^{8s''/(1+4s'')}, \sqrt{n}\sigma^2) = \sqrt{n}\sigma^2$ si et seulement si $n < n(\sigma) = 2^{J(\sigma)}$ avec

$$J(\sigma, s'', R) = \frac{1}{s'' + 1/4} \log_2\left(\frac{R}{\sigma} \right). \tag{2.23}$$

Dans ce cas la, vitesse de séparation minimax est identique au deuxième exemple (celui de la sous-section 2.3.4) lorsque $k = n$. Par conséquent, l'espace des alternatives comprend des vecteurs qui ne sont pas creux. Il n'y a rien à gagner en effectuant une réduction de dimension. Cette condition remplace d'une certaine manière la méconnaissance de k par la méconnaissance de s et p, cependant elle permet de donner de manière précise les cas dans lesquels la réduction de dimension n'est pas pertinente.

Chapitre 3

Tests minimax par seuillage

> Quand nous creusons dans la vérité pour la pénétrer, elle creuse aussi en nous pour prendre possession de nos âmes.
>
> Biran

Dans ce chapitre, nous donnons une autre explication à la nécessité de la réduction de dimension que celle donnée au chapitre précédent. Nous décrivons en détails les tests par seuillage. Nous discutons rapidement le choix du seuil, et effectuons une étude numérique.

3.1 Introduction

Nous allons construire un test sur la norme l^2 et expliquer l'intérêt d'une procédure de réduction de dimension pour ce type de test. La procédure de réduction de dimension que nous allons décrire a été donnée par deux personnes distinctes : Fan [30],[31] et Spokoiny [68]. Les travaux de Fan ne portent pas sur la théorie minimax. Le fond des remarques qui sont faites dans ce chapitre n'a donc rien d'original, ce chapitre nous permet d'introduire et de justifier l'intérêt d'un certain nombre de procédures que nous allons utiliser dans la suite de ce mémoire (en particulier dans le Chapitre 3 de la Partie 3, mais aussi dans la Partie 2).

Dans tout ce chapitre, nous supposons observer les n premières coordonnées d'une suite $\theta \in l^2$ dans du bruit :

$$Y_j = \theta_j + \sigma\xi_j, \quad \xi_j \rightsquigarrow \mathcal{N}(0,1), \quad j = 1, \dots, n, \tag{3.1}$$

où $\sigma > 0$ est supposé connu. Nous allons rappeler la construction d'une procédure de test basée sur une réduction de dimension pour tester des hypothèses du type

$$H_0 \ : \ \|\theta\|_{\mathbb{R}^n} = 0 \text{ contre } H_1 \ \ \theta \in \mathcal{F}(R) \cap \{\eta \in l^2 \ \ \|(\eta_1, \dots, \eta_n)\|_{\mathbb{R}^n} \geq \rho\}, \tag{3.2}$$

où $\mathcal{F}(R)$ est une partie de l^2 que nous préciserons par la suite et ρ est un paramètre à déterminer pour garantir une puissance satisfaisante (cf chapitre précédent). Cette distance ρ dépend de $\mathcal{F}(R)$, de n, du niveau α et de la puissance β que l'on souhaite obtenir.

3.2 Problématique : le test sans seuillage

Nous allons voir dans quel cas le test consistant à examiner la norme l^2 des observations n'est pas performant. Cela apportera une autre explication à la nécessité d'une réduction de dimension. Ce point de vu est celui de Fan [30]. Il est naturel, pour tester les hypothèses définies par (3.2), de chercher à mesurer la norme l^2 de θ^n à partir des observations $(Y_i)_{i=1,\dots,n}$. Le test suivant repose sur une mesure de la norme l^2 des observations :

$$\psi = 1_{R_\alpha}, \quad R_\alpha = \{T \geq v_{1-\alpha}\}, \tag{3.3}$$

où $v_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de T sous H_0, et T vaut :

$$T = \frac{1/n\sum_{j=1}^n Y_j^2 - \mathbb{E}_0[1/n\sum_{i=1}^n Y_i^2]}{\sqrt{Var_0(1/n\sum_{i=1}^n Y_i^2)}} = \frac{\sum_{j=1}^n (Y_j^2-\sigma^2)}{\sqrt{2n\sigma^4}}.$$

La statistique T suit asymptotiquement une loi normale centrée réduite sous H_0 et il est clair que si l'on remplace $v_{1-\alpha}$ par $z_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite, le niveau du test est asymptotiquement α. On remarque expérimentalement que lorsque $n \approx 100$ et que $\alpha \approx 0.01$, l'utilisation de l'approximation normale amène à effectuer un test à un niveau à peu près égal à 2α, aussi, pour ces valeurs de n (ou des valeurs plus petites), il est préférable de calculer $v_{1-\alpha}$ grâce aux quantiles du χ^2 à n degrés de liberté.

La dimensionalité du problème, qui correspond au nombre de termes présents dans la somme définissant T et permettant de mesurer l'énergie du signal (la norme l^2 des observations), réduit la puissance de ce type de test. C'est ce qu'illustre la proposition suivante :

Proposition 3.1. *[Fan [31]] Avec les notations introduites précédemment, on a :*

$$P_\theta(T \leq v_{1-\alpha}) = P\left(Z \leq \frac{v_{1-\alpha}}{\sqrt{1+2\|\theta\|^2/(n\sigma^2)}} - \frac{\|\theta\|_2^2}{\sqrt{2n\sigma^4+4\sigma^2\|\theta\|_{\mathbb{R}^n}^2}}\right), \tag{3.4}$$

où Z est une somme de n variables aléatoires indépendantes identiquement distribuées telle que $\mathbb{E}[Z]=0$, $Var(Z)=1$.

Souvent, θ ne contient que quelques coefficients de forte amplitude. La puissance est donc diminuée par la présence dans (3.4) du terme $1/\sqrt{2n\sigma^2+4\sigma^2\|\theta\|_{\mathbb{R}^n}^2}$ alors que peu de coefficients parmi les n présents contribuent à augmenter $\sum_{i=1}^n \theta_i^2 = \|\theta\|_{\mathbb{R}^n}$. Ce manque à gagner indique l'utilité d'une procédure de réduction de dimension. En d'autres termes, il est naturel et souhaitable de chercher à sélectionner $I \subseteq \{1 \cdots n\}$ tel que

$$\sum_{j\in I}\theta_j^2 \approx \sum_{j=1}^n \theta_j^2 \;\; et \;\; |I| \ll n.$$

Les méthodes de sélection déjà envisagées dans la littérature sont multiples (voire par exemple ([30])), nous allons maintenant présenter l'une d'entre elle.

3.3 Test par seuillage et puissance du test

Le test que nous présentons ici sera utilisé au Chapitre 3 de la partie III et a été introduit par Fan [31] et Spokoiny [68].

3.3.1 Construction du test.

Le test par seuillage correspond à rechercher les coefficients significativement grands, puis à éliminer les autres coefficients (trop proches de 0) avant d'évaluer l'énergie du signal. Le seuil utilisé peut être plus ou moins grand dans différentes parties du signal. En posant $\lambda_n = (\lambda_{in})_{i=1,\dots,n}$, l'ensemble des indices associés aux coefficients significatifs est

$$\hat{I}(n, \lambda_n) = \{i \in \{1, \dots, n\} \ : \ |Y_i| \geq \lambda_i\}, \tag{3.5}$$

et

$$\mathcal{E}_{\lambda_n} = \sum_{i \in \hat{I}(n,\lambda_n)} Y_i^2$$

est l'énergie de la partie significative du signal. Nous reviendrons par la suite au choix du vecteur de seuils λ_n. La statistique de test est alors obtenue en centrant et en normalisant la fonctionnelle d'énergie construite :

$$S_{\lambda_n} = \frac{\mathcal{E}_{\lambda_n} - \mathbb{E}_0\left[\mathcal{E}_{\lambda_n}\right]}{\sqrt{Var_0(\mathcal{E}_{\lambda_n})}}, \tag{3.6}$$

où $\mathbb{E}_0$ et Var_0 sont respectivement l'espérance et la variance calculées sous H_0 (sous la loi P_0). La discussion concernant le calcul de ces expressions est reportée plus bas. On a alors la proposition suivante

Proposition 3.2 (Spokoiny). *Soit $u \in \mathbb{R}$. Si $\sum_{i=1}^n \theta_i^2 1_{|\theta_i| \geq \sigma\lambda_{in}} \geq \sqrt{Var_0(\mathcal{E}_\lambda)}u$, alors on a :*

$$P_\theta(S_{\lambda_n} < u) = P\left(Z_n \geq B(n, \theta, \lambda_n)\right)$$

où

$$B(n, \theta, \lambda_n) \geq \frac{1}{2} \frac{\sum_{i=1}^n \theta_i^2 1_{|\theta_i| \geq \sigma\lambda_{in}} - \sqrt{Var_0(\mathcal{E}_\lambda)}u}{\left(4\sigma^2 \|\theta\|^2_{\mathbb{R}^n} + 2\sigma^4 \sum_{i=1}^n (1_{|\theta_i| \geq \sigma\lambda_{in}/2} + \lambda_{in}^4/2e^{-\frac{\lambda_{in}^2}{2\sigma^2}})\right)^{1/2}},$$

Z_n est une somme de n variables aléatoires (V_{i,λ_n}) telle que $E[Z_n] = 0$ $var(Z_n) = 1$.

Cette proposition résulte directement des lemmes 6.3 et 6.4 de Spokoiny ([68]). Nous ne donnons pas ici les arguments permettant de choisir le seuil $u(\alpha)$ pour assurer un niveau α au test. Sous certaines hypothèses sur λ_n la variable aléatoire Z_n de la proposition est asymptotiquement normale, nous renvoyons le lecteur à l'article de Spokoiny pour le seuil que nous utiliserons. Notons seulement que du fait que les variables aléatoires (V_{i,λ_n}) dépendent de n, cette normalité asymptotique n'est pas classique et il faut, pour l'obtenir, appliquer un théorème limite de type Lindenberg Feller (voir [56]).

3.3.2 Interprétation de la proposition précédente

La Proposition 3.2 nous donne un renseignement intéressant sur le lien entre le choix du seuil et la puissance du test. Tout d'abord, notons que la puissance du test est bonne lorsque $B(n,\theta,\lambda)$ est grand. La pire des valeur de $B(n,\theta,\lambda)$ sur la classe $\mathcal{F}(R)\cap\{\eta\in l^2 \quad \|(\eta_1,\dots,\eta_n)\|_2\geq\rho\}$ (ensemble des paramètres des distributions alternatives) détermine donc une borne supérieure du risque minimax.

Pour comprendre les conséquences de la proposition précédente, supposons que $\lambda_n=(\lambda_{1n},\dots,\lambda_{nn})$, que θ a k coordonnées non-nulles, et que si i est une de ces coordonnées $|\theta_i|\geq\sigma\lambda_{in}$. Supposons aussi que

$$\sum\lambda_{in}^4 e^{-\frac{\lambda_{in}^2}{2\sigma^2}}\leq kc_2,$$

pour c_2 une constante positive. Dans ce cas, on a :

$$B(n,\theta,\lambda_n)\geq\frac{\|\theta\|_2-\sqrt{Var_0(\mathcal{E}_{\lambda_n})}u}{\sqrt{\sigma^2R^2+k(c_2+2)\sigma^4}}.$$

Ainsi, la puissance obtenue est, à une constante indépendante de n près, comparable à celle que l'on aurait si l'on connaissait exactement les k coordonnées non nulles de θ (comparer l'équation précédente avec l'équation (3.4)). En ce sens l'inégalité de la proposition précédente peut être vue comme une inégalité oracle.

3.3.3 Calcul de l'espérance et de la variance de $\mathcal{E}_\lambda$ sous H_0,

Pour ce calcul, il suffit de savoir calculer $b_2(\lambda)=\mathbb{E}[\xi^2 1_{|\xi|\geq\lambda}]$ et $b_4(\lambda)=\mathbb{E}[\xi^4 1_{|\xi|\geq\lambda}]$. Pour le calcul de ces quantités, Fan dans [30] obtient (par quelques intégrations par partie) la relation

$$b_k(\lambda)=\sqrt{2/\pi}\lambda(1+\lambda^{-2}+O(\lambda^{-4}))e^{-\lambda^2/2}$$

Abramovich et al, dans [4] affirment qu'une meilleure expression (surtout pour les petites valeurs de n) est donnée par un développement de Laurent : pour $\lambda>1$

$$b_4(\lambda)=3-\sqrt{2/\pi}\Lambda^5/5+\Lambda^7/(7\sqrt{2\pi})+o(\Lambda^8)$$

où $\Lambda=\min(\lambda,1/\lambda)$.

3.3.4 Test minimax sur les boules de Besov

Soient $q>0$, $p>0$ et $s'>max((1/p-1/q),0)$. Nous allons nous placer dans le cas où $\mathcal{F}=\mathcal{B}_{s',p,p}(R)$. Nous rappelons la définition des ellipsoides de Besov :

$$\mathcal{B}_{s',p,q}(R)=\left\{f\in l^2,\ \sum_{j\geq0}\left[2^{js}\left(\sum_{k=2^j}^{2^{j+1}-1}|f_k|^p\right)^{1/p}\right]^q\leq R^q\right\}=\left\{f\in l^2,\ \ \|f\|_{b_{s',p,q}}\leq R\right\},\quad(3.7)$$

ce type de boule est lié à l'image d'une boule de Besov (espace de fonction, voir Annexe A) par la transformée en ondelette. Soit $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une gaussienne centrée réduite, et

$$\lambda_{in}=\begin{cases}4\sqrt{8\log(2^{j-J(\sigma,s'',R)})} & \text{si } i\in[2^j;2^{j+1}-1] \quad\text{et}\quad j\geq J(\sigma,s'',R),\\ 0 & \text{sinon}\end{cases},\qquad(3.8)$$

où $J(\sigma, s'', R)$ est défini par (2.23). Soit enfin le test

$$\psi_{\sigma,s'',R} = 1_{S_{(\lambda_{in})} > z_{1-\alpha}}. \tag{3.9}$$

Ce test est utilisé Chapitre 3 Partie III.

Il est asymptotiquement de niveau α et permet d'atteindre la vitesse minimax asymptotique pour tester les hypothèses définies par (3.2) et $\mathcal{F} = \mathcal{B}_{s,p,q}(R)$ (cf Spokoiny ([68])). En d'autres termes il existe une constante c positive telle que lorsque σ tend vers 0 :

$$\alpha(\psi^*) \leq \alpha_0 + o(1) \text{ et } \beta\left(\psi^*, \mathcal{B}_{s',p,q}(R) \cap \{\eta \in l^2 \ \ \|(\eta_1, \ldots, \eta_n)\|_2 \geq c\rho\}\right) \leq \beta_0 + o_\sigma(1) \tag{3.10}$$

où

$$\rho = \rho(\mathcal{B}_{s,p,q}(R), \alpha_0 + o_\sigma(1), \beta_0 + o_\sigma(1), \sigma) = R^{2/(1+4s'')}\sigma^{8s''/(1+4s'')}.$$

Remarque 3.1. *La valeur de seuil λ_{in} donnée par (3.8) dépend uniquement de l'intervalle dyadique auquel i appartient. On parle de seuil dépendant de l'échelle. Cette dénomination prend sens dans une base d'ondelette, dans laquelle les coefficients associés aux indices $i \in [2^j; 2^{j+1} - 1]$ sont les coefficients reflétant une même échelle de l'analyse multi-résolution (voir Annexe A). L'idée d'un seuil dépendant de l'échelle est due à Delyon et Iouditski [22] (ceux-ci ne l'utilisent pas pour des tests).*

Remarque 3.2. *Le test défini par (3.9) dépend de paramètres inconnus a priori : σ, R et s''. Nous donnerons une méthode pour estimer σ et nous supposerons que σ est connu. Spokoiny [68] propose un algorithme adaptatif pour effectuer un test ne nécessitant pas de connaître ces paramètres a priori. Dans la pratique, pour des spectres de taille comprise entre 256 et 2048, il nous a semblé que le choix $J(\sigma, s'', R) = 3$ offrait un bon compromis. La méthode de test par sélection de modèle décrite dans le paragraphe suivant est une méthode adaptative.*

3.4 Une alternative au seuillage : la sélection de modèle

Le test par seuillage présenté est un test qui s'effectue en deux temps : le premier consiste en la sélection d'un sous espace matérialisé par un certain nombre de coefficients, le deuxième consiste en la mesure de l'énergie du signal dans ce sous-espace. Ceci constitue une procédure de sélection de modèle. Dans le cadre de la sélection de modèle une procédure alternative aux procédure de type seuillage a été donnée par Baraud et Al. ([8]). Dans notre cadre, ce test est le suivant. Soit $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ une famille finie, indexée par $\mathcal{M}$, de sous espaces de $\mathbb{R}^n$ de dimension $(D_m)_{m\in\mathcal{M}}$ avec $n > D_m > 0$. On suppose ici que $n \geq 2$. Etant donné $\alpha \in]0, 1[$, soit

$$T_{\mathcal{S},\alpha} = \sup_{m\in\mathcal{M}} \left\{ \frac{(n - D_m)\|\Pi_m Y\|^2_{\mathbb{R}^n}}{D_m\|Y - \Pi_m Y\|^2_{\mathbb{R}^n}} - \bar{F}^{-1}_{D_m, n-D_m}(\alpha_m) \right\},$$

où $\{\alpha_m, m \in \mathcal{M}\}$ est une suite de nombres dans $]0, 1[$ tels que

$$\sum_{m\in\mathcal{M}} \alpha_m \leq \alpha.$$

D'après le théorème de Bonferroni (cf chapitre suivant théorème 4.1) le test

$$\phi_{\mathcal{M}} = 1_{T_\alpha > 0} \tag{3.11}$$

est de niveau inférieur ou égal à α.

Remarque 3.3. *Un des intérêts de ce test réside dans le fait que le niveau du test n'est pas déterminé de manière asymptotique. Baraud et Al. ([8]) proposent aussi de déterminer α_m par simulation afin d'assurer au test un niveau exactement égal à α. Nous n'avons pas voulu utiliser cette stratégie car nous voulons (pour la troisième partie de ce mémoire) un test pour lequel la relation entre α et α_m est simple et rapide à calculer (pour des valeurs de α pouvant varier).*

Ce test peut être adapté à un grand nombre de situations. Il s'agit seulement de choisir soigneusement la famille $\mathcal{S}$ et la suite de nombre $(\alpha_m)_m$. Baraud et Al. ([8]) obtiennent des vitesses minimax non asymptotiques dans un certain nombre de cas standards et donnent des choix de $\mathcal{S}$ et $(\alpha_m)_m$ qui permettent de les atteindre. Nous rappelons ce choix dans le cas ou l'alternative a tester est du type de celles étudiées précédemment (voir ([8]) pour plus de détails). Soit $(e_i)_{i=1,\dots,n}$ la base canonique de $\mathbb{R}^n$. Pour tout $k \in \mathcal{M}_1 = \{2^j,\ j \geq 0\} \cap \{1,\dots,[n/2]$, soit $S_{(k,1)}$ le sous espace engendré par $(e_1,\dots,e_k)$ et pour tout $k \in \mathcal{M}_2 = \{1,\dots,n\}$ soit $S_{(k,2)}$ la droite vectorielle engendrée par e_k. La famille de modèles que nous allons utiliser est finalement indexée par $\mathcal{M} = \{(k,1), k \in \mathcal{M}_1\} \cup \{(k,2), k \in \mathcal{M}_2\}$ et donnée par $(S_{(i,j)})_{(i,j)\in\mathcal{M}}$. Définissons maintenant $(\alpha_m)_{m\in\mathcal{M}}$. Soit $k_0 = \sup \mathcal{M}_1$ et $\alpha > 0$. Pour tout $k \in \mathcal{M}_1$ et $k \neq k_0$, nous avons posé $\alpha_{(k,1)} = \frac{\alpha}{4|\mathcal{M}_1|}$, $\alpha_{(k_0,1)} = \frac{\alpha}{4}$. Pour tout $k \in \mathcal{M}_2$, soit $\alpha_{(k,2)} = \frac{\alpha}{2n}$.

3.5 Etude comparative

Afin de choisir une méthode de test efficace dans l'utilisation que nous en ferons au Chapitre 3 Partie III, nous avons étudié les tests décrits dans le cas où l'alternative inconnue à détecter est une lorentzienne définie pour $a, m \in \mathbb{R}$ par

$$g_{a,m}(x) = \frac{a}{1 + a^2(x-m)^2}.$$

Ce type de courbe est en théorie ce que les équations physiques la spectroscopie par résonance magnétique prévoient en un pic donné. Nous avons choisi $a = 0.2$, $m = 1048$, $n = 2048$, et six approches différentes pour tester les hypothèses correspondantes. Dans les deux premières, on observe directement

$$Y_i = g_{0.2,1048}(i) + \sigma\epsilon_i, \tag{3.12}$$

et dans les trois dernières, on effectue une transformée en ondelettes discrète $\mathcal{W}$ (voir le livre de Mallat [53]) sur $(Y_i)_i$ et on observe donc

$$Z_i = (\mathcal{W} g_{0.2,1048})_i + \sigma\eta_i. \tag{3.13}$$

Dans tous les cas $(\epsilon_i)_{i=1,\dots,2048}$ et $(\eta_i)_{i=1,\dots,2048}$ sont des variables aléatoires gaussiennes centrées réduites indépendantes et σ est un réel positif que nous faisons varier pour faire évoluer le rapport signal sur bruit (la Figure 3.1 illustre les différentes valeurs prises).

Le premier test est un test du χ^2, celui défini par (3.3) à partir des observations (3.12) (la puissance correspondante est la courbe « sans seuillage » dans la Figure 3.1).

Le deuxième est un test du χ^2 avec seuillage universel : les observations sont données par (3.12), le test est ψ_{λ_U} où ψ_λ est donné par (3.9) et $\lambda_U = \sigma\sqrt{2\log(n)}$ (la puissance correspondante est la courbe « seuillage » dans la Figure 3.1).

Les quatres derniers tests sont construits à partir des observations données par (3.13). Le test marqué U sur la Figure 3.1 est ψ_{λ_U} où ψ_λ est donné par (3.9) et $\lambda_U = \sigma\sqrt{2\log(n)}$. Le test

marqué FDR dans la Figure 3.1 est $\psi_{\lambda_{FDR}}$ où ψ_λ est donné par (3.9) et λ_{FDR} est obtenu comme suit. Comparer les $(|Z|_{(k)})_k$ (Y_i ordonnées par ordre de valeurs absolues décroissantes) à la queue de distribution d'une gaussienne $t_k = \sigma z\left(\frac{qk}{2n}\right)$:

$$k_{fdr} = max\left\{k : \ |Z|_{(k)} \geq t_k\right\}$$

et définir :

$$\lambda_{FDR} = t_{k_{fdr}}.$$

Le test marqué DY dans la Figure 3.1 est le test défini par l'équation (3.8). Le quatrième test effectué dans la base d'ondelette est le test de Baraud et Al. ([8]). Ce test a été présenté à la sous-section précédente.

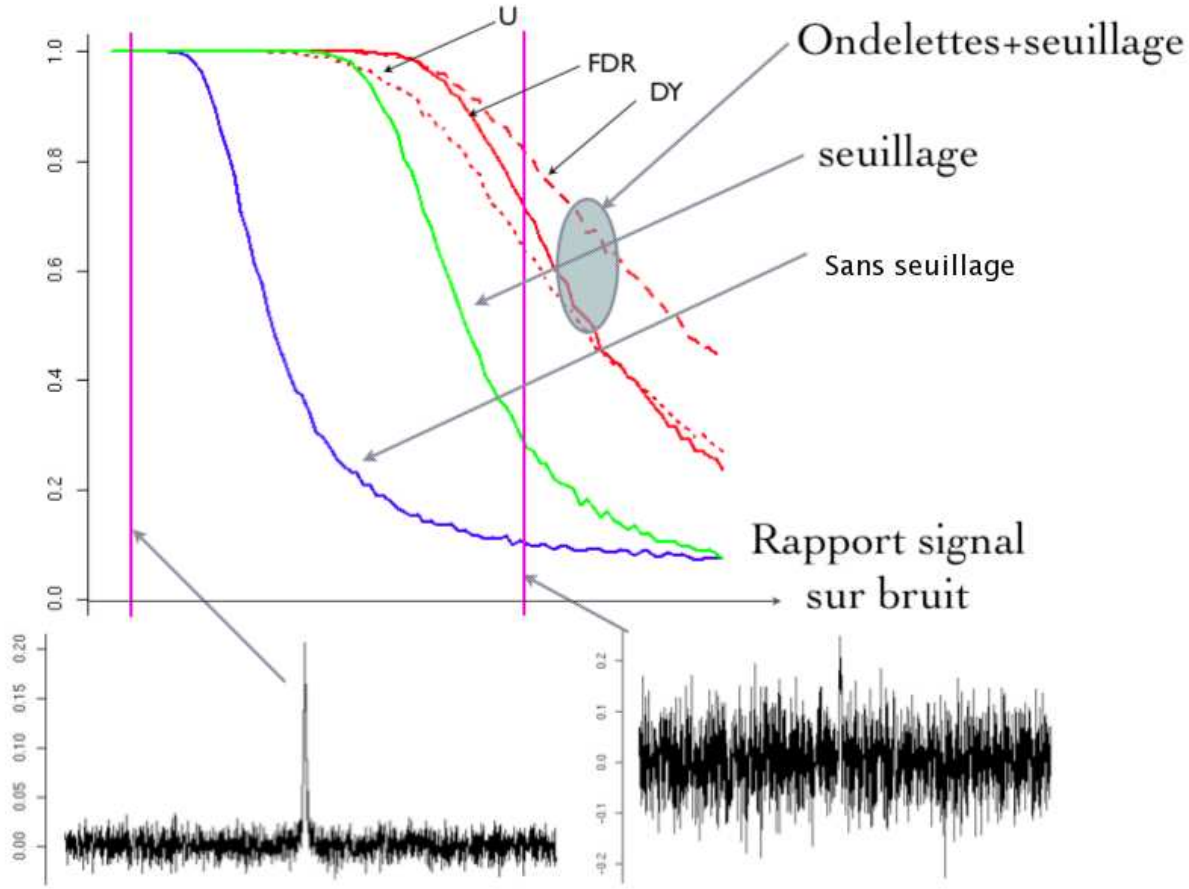


FIG. 3.1 – Puissances des tests en fonction du rapport signal sur bruit. U : seuil universel $\lambda = \sqrt{2\log(n)}$; FDR : seuil FDR ; DY : seuil Delyon Youditski

On peut noter sur la Figure 3.1 que l'effet du seuillage est bénéfique (différence de puissance entre « sans seuillage » et « seuillage »). La transformée en ondelette a pour effet d'augmenter la puissance du test par seuillage (comparer « seuillage » et « U » dans la Figure 3.1). Des trois types de seuillage envisagés dans la base d'ondelette, le seuillage dépendant de l'échelle semble

être celui qui amène le test le plus puissant. Cette étude est à mettre en parallèle avec le caractère optimal du test avec seuillage dépendant de l'échelle. Bien évidement puisque dans notre étude pratique nous ne regardons qu'une alternative fixée, ce parallèle reste incomplet. Notons enfin que la Figure 3.2 semble nous indiquer que les performances du test par seuillage sont meilleurs que celles du test par sélection de modèle. Il ne faut cependant relativiser ces différences de performances puisque le test par sélection de modèle est adaptatif alors que pour le test par seuillage, nous avons choisi de fixer une régularité a priori. Notons pour finir que la principale perspective dans ce type de tests est de parvenir à traiter des cas où l'on observe

$$Y_i = \theta_i + \sigma_i \xi_i, \quad i = 1, \dots, n$$

où $(\sigma_i)_{i=1,\dots,n}$ aussi est inconnu et où l'on veut tester des hypothèses du type de celles définies par (3.2).

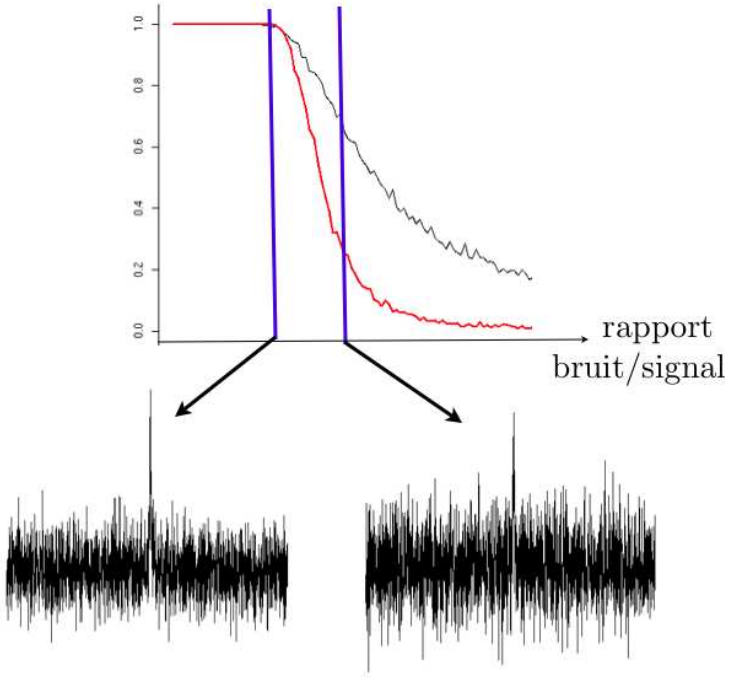


FIG. 3.2 – Puissances des tests par seuillage (seuil Delyon Youditski, courbe en noire) et par sélection de modèle (en rouge)

Chapitre 4

Tests multiples

> Je raconterai cette histoire en toute honnêteté ; je parviendrai peut-être ainsi à la comprendre moi-même.
>
> J.L. BORGES, Guayaquil

Dans ce chapitre, nous introduisons la problématique des tests d'hypothèses multiples. Nous définissons la proportion de rejet à tort (FDR) et le FWER. Nous donnons des procédure pour contrôler ces deux types d'erreur. Nous terminons en présentant l'application des procédures de test multiple à une méthode d'estimation par seuillage donnée par Abramovich et ses collaborateurs [2].

4.1 Problématique des tests multiples

La problématique des tests multiples est la suivante. On cherche à effectuer plusieurs tests et à contrôler non pas l'erreur commise pour chacun d'eux séparément, mais une erreur globale. On peut par exemple considérer la probabilité de faire une erreur de première espèce ou d'avoir une proportion d'erreur de première espèce d'un certain ordre. Si l'on décrit les hypothèses par

$$H_0^i \quad contre \quad H_1^i \quad i = 1, \ldots, m, \tag{4.1}$$

les performances jointes de tests de ces hypothèses peuvent être décrites par le tableau ci dessous, qui compte les rejets et acceptations, à tort ou à raison. (on a choisi m_0 fois H_0, m_1 fois H_1 ...).

	accepté	rejeté	total
$H_0 = 0$	U	V	m_0
$H_0 = 1$	T	S	m_1
total	W	R	m

Plusieurs types de contrôles ont été envisagés. Nous ne cherchons pas ici à faire une description exhaustive de tout ce qui a été fait, tant il est vrai que les tests multiples constituent un vaste domaine des statistiques. Nous allons introduire et motiver des procédures qui sont liées à nos problèmes et dont nous ferons usage dans la suite. Les deux erreurs les plus couramment étudiées sont :

- le contrôle de la probabilité qu'au moins une erreur de première espèce advienne $FWER = P(V \geq 1)$ (Familly wise error). La stratégie la plus simple pour s'assurer que $FWER \leq \alpha$ est d'utiliser la procédure de Bonféronni qui consiste à faire en sorte que l'erreur de première espèce associée à chacune des hypothèses soit bornée par α/n. Même s'il existe un certain nombre de méthodes plus performantes (moins conservatrices) que celle de Bonferonni, contrôler le $FWER$ n'est pas toujours ce qu'il y a de plus adapté à nos problèmes.
- le contrôle de l'espérance de la proportion de rejet à tort : le FDR (False Discovery Rate) :

$$FDR = E[Q] \quad \text{où} \quad Q = \begin{cases} V/R & si \quad R > 0 \\ 0 & si \quad R \leq 0 \end{cases}.$$

  La méthode la plus simple et efficace pour le contrôle du FDR a été introduite par Benjamini et Hochberg [9]. Nous la décrivons dans la suite.

Les hypothèses multiples qui nous intéresseront. Nous allons maintenant présenter l'exemple que nous utiliserons dans la suite. Soit $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2 I_m)$ un vecteur gaussien de $\mathbb{R}^m$. On cherche à tester simultanément les hypothèses :

$$H_{0i} \ : \ \mu_i = 0 \ \text{ contre } \ H_{1i} \ : \ \mu_i \neq 0. \tag{4.2}$$

Pour chaque couple d'hypothèses, le test de Neymann Pearson est $\psi_\alpha^i(X_i) = |X_i| > \sigma z_{1-\alpha/2}$. Avant de donner les méthodes standards du contrôle de FWER et FDR, rappelons la définition de la p-valeur.

p-valeur ou seuil critique du test. Soit un tests $\psi_\alpha(X)$ pour tester les hypothèses H_0 contre H_1, construit à partir d'une statistique X. Ce test peut être vu comme une fonction de α le niveau du test. La valeur $\alpha^*(X)$ est dite p-valeur ou seuil critique du test si

$$\psi_\alpha(X) = \begin{cases} 1 & si \quad \alpha > \alpha^*(X) \\ 0 & si \quad \alpha \leq \alpha^*(X) \end{cases}.$$

Dans la suite de ce chapitre, nous supposerons que $P(\alpha^*(X) \leq \alpha) \leq \alpha$ (c'est le cas pour le test des hypothèses définies par (4.2) et avec la statistique qui y est donnée). L'utilisation des p-valeurs est naturelle et courante chez tous les utilisateurs des statistiques. Les p-valeurs évitent à l'utilisateur de l'outil statistique d'aller comparer la valeur de la statistique de test à un seuil dans une table de loi ; ils sont à ce titre très utiles. Ainsi, si l'on veut tester deux hypothèses simples H_0 contre H_1 à un niveau α et que nous avons à disposition une p-valeur $\alpha^*(X)$ pour tester ces hypothèses. Il suffit alors de rejeter H_0 si $\alpha_i^*(X_i) \leq \alpha$. Dans le cadre d'un test multiple il existe une procédure simple si l'on cherche à contrôler le FWER. C'est la procédure de Bonferroni, donnée par le théorème suivant.

Theoreme 4.1 (Procédure de Bonferroni [9]). *Si pour $i = 1, \dots, m$, l'hypothèse H_{0i} est rejeté lorsque $\alpha_i^*(X_i) \leq \alpha/m$, alors $FWER \leq \alpha$.*

Démonstration. Notons I l'ensembles des indices i tels que H_{0i} soit vrai et $|I|$ le cardinal de I. Ainsi

$$FWER \leq \sum_{i \in I} P(\alpha_i^*(X_i) \leq \alpha/m) \leq \frac{|I|}{m}\alpha \leq \alpha.$$

□

Notons que dans cette inégalité l'imperfection de la borne est surtout due au fait que $\frac{|I|}{m}$ peut être très petit si $|I|$ est petit par rapport à m. Ceci n'est bien sûr pas sans rappeler le problème concernant la puissance du test du χ^2 lorsque dans l'alternative peu de coefficients sont non nuls. C'est bien en cela que les solutions données à ce problème nous intéresserons. La première solution à ce problème a été donnée par Holm [42], et un grand nombre de différentes procédures ont suivi. Ces procédures reposent sur l'heuristique suivante. Si l'on effectue les m tests de manière séquentielle mais en utilisant des statistiques indépendantes, et qu'après k tests, aucun rejet n'a été effectué il reste alors $m - k$ tests à effectuer sur lesquels le « budget » d'erreur est toujours de α. On peut donc sur ces tests être plus audacieux. La procédure de Holm est la suivante :

- **Initialisation** : Ordonner les p-valeurs associées aux m tests considérés : $p_{(1)} \leq \cdots \leq p_{(m)}$. Poser $k = 1$.
- **Itération** : Si $p_{(k)} \geq \frac{\alpha}{m-k+1}$ accepter $H_{0k}, \dots, H_{0m}$ et s'arrêter. Sinon, rejeter H_{0k}, faire $k = k + 1$ et refaire l'étape d'itération.

Theoreme 4.2 (Holm [42]). *La procédure de Holm permet d'avoir $FWER \leq \alpha$.*

Démonstration. Supposons que I est l'ensemble des $i \in \{1, \dots, m\}$ tels que H_{0i} est vrai. Soit

$$j = \text{argmin}\{k : p_{(k)} = \min_{i \in I} p_i\},$$

pour cet entier, on a $j \leq m - |I| + 1$.

La procédure de Holm effectue un rejet à tort si

$$\forall k \leq j \quad p_{(k)} \leq \frac{\alpha}{m - k + 1},$$

et dans ce cas

$$\min_{i \in I} p_i = p_{(j)} \leq \frac{\alpha}{m - j + 1} \leq \frac{\alpha}{|I|}.$$

Ainsi,

$$FWER \leq P\left(\min_{i \in I} p_i \leq \frac{\alpha}{|I|}\right),$$

et du fait de la sous-additivité des probabilités, on a :

$$FWER \leq \sum_{i \in I} P\left(p_i \leq \frac{\alpha}{|I|}\right) \leq \alpha.$$

□

4.2 Méhodes de Benjamini, Hochberg, Yekutieli et Storey

La procédure de Benjamini et Hochberg ne vise pas à contrôler le FWER, mais le FDR. Cependant elle repose sur les mêmes idées que la procédure de Holm.

Première méthode. La méthode introduite par Bonferoni et Hochberg est la suivante. Les m tests fournissent m p-valeurs $p_i = \alpha_i^*(X_i)$ que l'on classe par ordre croissant $(p_{(i)})_{i=1,\dots,m}$. On cherche ensuite

$$k_{fdr} = \max_{1 \leq ß \leq m} \left\{ i \ : \ p_{(i)} \leq \frac{i}{m} q^* \right\}. \tag{4.3}$$

Si un tel k n'existe pas, H_{0i} est accepté pour tout $i \in \{0, \dots, m\}$, sinon, les hypothèses correspondants aux k plus grandes p-valeurs p^i sont acceptées, les autres sont rejetées. Cette procédure permet de contrôler le FDR. Notons juste qu'il est facile de voir que $FDR \leq FWER$ et que lorsque le nombre d'hypothèses H_{0i} fausses est différent de 0, cette inégalité est souvent stricte. Aussi une méthode visant à contrôler le FDR sera forcement moins conservatrice (plus libérale) en ce sens qu'elle rejettera plus facilement H_{0i}. Bonferoni et Hochberg obtiennent pour leur algorithme, le résultat suivant :

Theoreme 4.3 (Benjamini,Hochberg [9])**.** *Si les statistiques de tests sont indépendantes la procédure décrite ci-dessus contrôle le FDR au niveau q^* :*

$$FDR = E[Q] \leq q^*$$

La démonstration se fait par récurrence et use essentiellement du fait que sous H_0 les p-valeurs sont des variables aléatoires identiquement distribuées de loi $\mathcal{U}(0,1)$.

Méthode dans le cas dépendant. Dans le cas de données dépendantes, c'est-à-dire si les observations X_i associées à chaque paire d'hypothèses $(H_{0i}, H_{1i}\ )$ sont dépendantes, Benjamini et Yekutieli [10] parviennent à montrer que le contrôle du FDR reste possible. Si $X = (X_1, \dots, X_n) \in \mathbb{R}^d$ vérifie la propriété PRDS (positive regression dépendancy) sur I_0 l'ensemble des hypothèses nulles vraies ; c'est-à-dire si pour toute partie D de R^d, et pour tout $i \in I_0$, $P(X \in D | X_i = x)$ est croissant avec x ; alors la procédure de Benjamini et Hochberg permet encore un contrôle du FDR.

Dans tous les cas il est possible de contrôler le FDR en appliquant la procédure de Benjamini et Hochberg en remplaçant q par $q/(\sum_{i=1}^m \frac{1}{i})$.

Idée de Storey pour augmenter la puissance. Dans tous les cas, la procédure de Benjamini et Hochberg s'avère trop conservatrice. Elle a connu de nombreuses déclinaisons améliorant plus ou moins ce problème. La procédure proposée par Storey [71] est à ma connaissance la plus simple et efficace. L'idée de Storey est que si l'on connaît un estimateur $\hat{\pi}_0$ supérieurement biaisé de la proportion π_0 de H_{0i} vrai, il est possible d'augmenter la puissance de la procédure de Benjamini et Hochberg en remplaçant $\frac{i}{m} q^*$ dans (4.3) par $\frac{i}{\hat{\pi}_0 m} q^*$.

Application à notre problème Pour tester les hypothèses définies par (4.2) les procédures de Benjamini et Hochberg ($FDR(X,q)$), ainsi que la procédure utilisant un estimateur de la proportion de H_{0i} vrai $SFDR(X, q, \hat{\pi}_0)$ sont les suivantes.

$$ALGO \;\; FDR(X, q, \sigma)$$
$$\underline{\text{Entrée}} : X = (X_j)_{j\in\{1,\dots,m\}}$$
$$\underline{\text{Sortie}} : \hat{I} \subset \{1, \dots, m\}$$

1. Classer les valeurs absolues des valeurs observées dans l'ordre décroissant : $|X|_{(i)}$
2. Les comparer à la queue de distribution d'une gaussienne :

$$t_i = z\left(\frac{qi}{2m}\right)$$

$$k_{fdr} = max(i : \;\; |X|_{(i)} \geq \sigma t_i).$$

3. Choisir le seuil de la manière suivante :

$$\lambda_{FDR} = t_{k_{fdr}}$$

4. L'ensemble d'indices recherché est alors

$$\hat{I} = \{i \in [1, m] \;\; tq \;\; |X_i| \geq \lambda_{FDR}\}.$$

$$ALGO \;\; SFDR(X, q, \hat{\pi}_0, \sigma)$$

$$\underline{\text{Entrée}} : X = (X_j)_{j\in\{1,\dots,m\}}, \hat{\pi}_0$$
$$\underline{\text{Sortie}} : \hat{I} \subset \{1, \dots, m\}$$

1. Classer les valeurs absolues des valeurs observées dans l'ordre décroissant : $|X|_{(i)}$
2. Les comparer à la queue de distribution d'une gaussienne :

$$t_i = z\left(\frac{qi}{2\hat{\pi}_0 m}\right)$$

$$k_{fdr} = max(i : \;\; |X|_{(i)} \geq \sigma t_i).$$

3. Choisir le seuil de la manière suivante :

$$\lambda_{FDR} = t_{k_{fdr}}$$

4. L'ensemble d'indices recherché est alors

$$\hat{I} = \{i \in [1, m] \;\; tq \;\; |S_i^1| \geq \lambda_{FDR}\}.$$

Nous utiliserons ces deux procédures dans le Chapitre 3 (segmentation d'image hyper-spectrale avec AWS) de la dernière partie de ce mémoire. Nous utiliserons la première dans le cadre de la classification de courbes pour la réduction de dimension. Dans ce cas, l'utilisation de cette procédure sera motivée par le résultat que nous allons exposer dans la sous-section suivante.

4.3 Une application aux méthodes d'estimation par seuillage

Nous présentons les méthodes d'estimation par seuillage dans l'Annexe A. Nous conseillons donc au lecteur qui ne serait pas familier de ces méthodes de lire l'Annexe A.

Soit $X \rightsquigarrow \mathcal{N}(\mu, \sigma_m^2 I_m)$ une variable aléatoire gaussienne de moyenne $\mu \in \mathbb{R}^m$ et de covariance diagonale égale à $\sigma_m^2 I_m$ ($\sigma_m > 0$). Soit la $i$eme coordonnée de l'estimateur par seuillage dur $\hat{\mu}^\lambda$ de μ donnée par

$$\hat{\mu}^\lambda(X)_i = X_i 1_{\{|X_i|>\lambda\}} = X_i 1_{\{i \in \hat{I}(\lambda)\}} \quad \text{où } \hat{I}(\lambda) = \{i \in \{1;\dots,m\} \text{ tq } |X_i| > \lambda\}.$$

L'estimateur de $\hat{\mu}^{FDR}$ par seuillage FDR est $\hat{\mu}^{\lambda_{FDR}}$. Le seuil λ_{FDR} est calculé par l'algorithme $FDR(X_i, q_m, \sigma_m)$ où $q_m \geq \gamma/\log(m)$ pour γ une constante positive et q_m a une limite $q \in [0, 1/2[$ quand m tend vers l'infini.

Nous rappelons que si $\Theta_m \subset \mathbb{R}^m$, le pire des risques l^2 d'un estimateur de $\hat{\mu}$ sur Θ_m est

$$\rho(\hat{\mu}, \Theta_m) = \sup_{\mu \in \Theta_m} E_\mu[\|\mu - \hat{\mu}\|_2^2]. \tag{4.4}$$

Le risque minimax d'estimation est alors le pire des risques obtenu pour le meilleur des estimateurs :

$$R_m(\Theta) = \inf_{\hat{\mu}} \rho(\hat{\mu}, \Theta_m). \tag{4.5}$$

Abramovich et ses collaborateurs [2] démontrent (Theoreme 1.1 de l'article) le théorème suivant :

Theoreme 4.4 (Abramovich, Benjamini, Donoho, Johnstone)**.** *Supposons que q_m a une limite $q < 1/2$ lorsque m tend vers l'infini et que $q_m \geq \gamma/\log(m)$ pour γ une constante positive. Soit $0 < p < 2$, $\Theta_m = l^p(m^{1/p}\eta_m/\sigma_m)$ avec*

$$l^p(R) = \{x \in \mathbb{R}^m \ : \ \sum_{i=1}^m \theta_i^p \leq R^p\}$$

et $\eta_m^q \in [m^{-1}\log^5(m), m^{-\delta}]$, $\delta > 0$. Alors, quand m tend vers l'infini,

$$\rho(\hat{\mu}, \Theta_m) = R_m(\Theta_m)\left\{1 + o_m(1)\right\}. \tag{4.6}$$

En d'autres termes le seuil FDR permet de construire un estimateur adaptatif de μ (minimax simultanément sur plusieurs types de boules l^p pour $p < 2$ et plusieurs risques).

Nous utiliserons ce résultat dans le Chapitre 2 de la Partie II, pour justifier une méthode de réduction de dimension en classification. La deuxième partie de ce mémoire fait le lien entre les problèmes d'estimation donnés ci-dessus et le problème de classification. Cela permet l'utilisation dans notre contexte du Théorème 4.4 dédié au problème de régression. Dans un cas voisin (plus complexe) du cas qui nous intéresse, c'est-à-dire du test des hypothèses définies par (4.2), Bunea et ses collaborateurs [17] montrent dans un certain cadre que si l'ensemble I_0 des indices associés aux coefficients non nuls de μ est de cardinal borné avec n et que l'on dispose d'un nombre assez grand d'observations de $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$, alors le nombre d'erreurs de première et seconde espèce

tend vers 0. Nous n'utiliserons pas ce type résultat, et il nous a semblé que le résultat obtenu par Abramovitch et al. relevait d'observations bien plus fortes. Nous ne voulons pas nous limiter au cas des boules de types l_0 (nombre de composantes non nulles borné), et nous aimerions connaître la vitesse à laquelle l'estimation des paramètres se fait. En effet, nous allons montrer (Partie II) que celle-ci nous donne la vitesse de convergence vers 0 de l'erreur d'apprentissage de notre procédure de classification.

Chapitre 5

Classification supervisée, Learning par plug-in

What I tell you three times is true.

Lewis Carroll *The Hunting of the Snark.*

Dans ce chapitre nous introduisons le problème de classification et plus particulièrement le problème de classification supervisée. Nous donnons pour ce problème la mesure de risque dans le cadre d'une approche bayesienne. Nous définissons l'excès de risque, et introduisons en la motivant une nouvelle mesure de risque : l'erreur d'apprentissage. Le théorème 5.2 que nous avons obtenu permet de faire le lien entre les deux mesures d'erreur dans le cas gaussien. Nous rappelons pour finir quelques résultats concernant la classification par plug-in.

5.1 Le problème de classification et plusieurs mesures d'erreur

La classification est le problème central de la théorie de l'apprentissage. Il consiste à prédire la nature y, appelée aussi classe ou label, d'une observation x. Dans le cas le plus simple, celui de la classification binaire, le label prend ses valeurs dans $\{0,1\}$, et dans le cas de la classification à K classes, y prend ses valeurs dans $\{1,\dots,K\}$. L'observation x est très souvent la réunion d'un certain nombre d'attributs numériques formant un vecteur de $\mathcal{X}=\mathbb{R}^p$, mais elle peut aussi être une courbe ou une image. La classification en dimension finie est le cas où $\mathcal{X}=\mathbb{R}^p$ et la classification de courbes est le cas où $\mathcal{X}$ est un espace fonctionnel de dimension infinie.

En classification à K classes, on construit une application g de $\mathcal{X}$ dans $\{1,\dots,K\}$ qui à une observation $x\in\mathcal{X}$ associe la prédiction faite. Cette application est une fonction de décision que l'on appelle classificateur. Ce classificateur se trompe sur l'observation x si $g(x)\neq y$.

Mesure d'erreur par un principe bayesien. Pour formaliser le problème d'apprentissage tel que nous l'envisageons, il faut introduire un formalisme probabiliste. Ainsi, nous supposons que (X, Y) est une variable aléatoire à valeur dans $\mathcal{X} \times \{1, \dots, K\}$ de loi P modélisant la distribution des observations et des classes associées. Si $k \in \{1, \dots, K\}$, nous notons P_k la loi de $X|Y = k$, cette loi modélise la distribution des observations issues de la classe k. On souhaite naturellement construire un classificateur performant, c'est-à-dire qui se trompe avec une probabilité la plus faible possible. Il existe dans ce problème K erreurs de natures différentes consistant à ne pas affecter une observation à la classe $k \in \{1, \dots, K\}$ alors que son label vaut effectivement k. Pour k fixé, l'erreur est mesurée par $P_k(g(X) \neq k)$. Il y a alors plusieurs démarches selon l'importance que l'on donne au divers types d'erreurs. La démarche bayesienne est la plus couramment utilisée. Elle consiste à mesurer l'erreur de classification grâce au risque bayesien : $P(g(X) \neq Y)$. Cette erreur peut être réécrite grâce à la formule de Bayes :

$$\mathcal{C}(g) = P(g(X) \neq Y) = \sum_{k=1}^{K} P(Y = k) P_k(g(X) \neq k). \tag{5.1}$$

Si les lois jointes et marginales de (X, Y) sont parfaitement connues, on peut calculer la règle de décision optimale aussi appelée règle de Bayes et notée g^*. Plaçons-nous dans le cas de la classification binaire ($K = 2$). On a

$$\inf_g \mathcal{C}(g) = g_2(P(Y = 0), P_1, P_0),$$

où $g_2(t, P_1, P_0)$ est définie au Chapitre 1 Partie I (cf équation 1.6). La règle de bayes est donnée (cf Chapitre 1 Partie I) par

$$g^*(x) = \begin{cases} 1 & \text{si } dP_1 > \frac{P(Y=0)}{P(Y=1)} dP_0 \\ c(P(Y=0)) & \text{si } dP_1 = \frac{P(Y=0)}{P(Y=1)} dP_0 \\ 0 & \text{si } dP_1 < \frac{P(Y=0)}{P(Y=1)} dP_0 \end{cases} \tag{5.2}$$

$$\text{où } c(P(Y=0)) = \begin{cases} 1 & \text{si } P(Y=0) > 1/2 \\ c \in [0,1] & \text{si } P(Y=0) = 1/2 \\ 0 & \text{si } P(Y=0) < 1/2 \end{cases}.$$

Si $\mathcal{X}$ est un espace d'état fini ou dénombrable la règle de bayes règle est :

$$g^*(x) = \begin{cases} 1 & \text{si } \eta(x) = P(Y = 1|X = x) > 1/2 \\ c(P(Y=0)) & \text{si } \eta(x) = \frac{1}{2} \\ 0 & \text{si } \eta(x) < 1/2 \end{cases} \tag{5.3}$$

La fonction $\eta(x) = P(Y = 1|X = x)$ est appelée fonction de régression. Dans le cas où $\mathcal{X}$ est quelconque, on peut définir $\eta(X)$ comme étant la variable aléatoire $P(Y = 1|X)$.

Remarque 5.1. *La constante c utilisée pour définir g^* n'a aucune influence sur le caractère optimal de la règle g^* associée. Ainsi, on ne définit pas une règle optimale, mais une famille de règles optimales indexées par le paramètre $c \in [0, 1]$. Dans la plupart des cas non dégénérés ($d_1(P_1, P_0) \neq 0$) rencontrés dans la pratique, deux éléments de cette famille diffèrent sur un ensemble de probabilité nulle. Par convention, et pour simplifier l'expression des règles utilisées, nous choisirons $c = 1$ dans toute la suite.*

A priori uniforme. On note que la mesure d'erreur (5.1) donne une grande importante à $P_k(g(X) \neq k)$ si la classe k a une forte probabilité d'apparition. En d'autres termes, dans le cadre bayesien, l'erreur $P_k(g(X) \neq k)$ associée à une affectation erronée d'une observation de la classe k, est d'autant plus importante que le label k apparaît avec une grande probabilité. Dans notre cadre, il n'est pas raisonnable de penser que la fréquence d'apparition d'une tumeur ayant un label $k \in \{1, \dots, K\}$ nous renseigne sur l'importance que l'on doit donner à l'erreur faite lorsque $g(X) \neq k$ sachant que Y vaut k. En effet, il n'est pas naturel d'attribuer à une tumeur apparaissant peu souvent un faible poids tant il est vrai que dans le domaine médical la rareté est souvent synonyme de pathologie sérieuse. Nous proposons donc de supposer que si j et k sont deux labels différents, les erreurs faites lorsque $g(X) \neq k$, sachant que Y vaut k, et lorsque $g(X) \neq j$, sachant que Y vaut j, sont d'égale importance. Ainsi, nous supposerons très souvent que $Y \rightsquigarrow \mathcal{U}(\mathcal{Y})$, ce qui implique

$$\mathcal{C}(g) = \frac{1}{K} \sum_{k=1}^{K} P_k(g(X) \neq k). \tag{5.4}$$

Plaçons nous dans le cas où $K = 2$ (classification binaire), $\mathcal{Y} = \{0, 1\}$, $Y \rightsquigarrow \mathcal{U}(\mathcal{Y})$ et $P_1 \sim P_0$. Dans ce cas, $\frac{P(Y=0)}{P(Y=1)} = 1$ et l'expression de g^* se simplifie. Il est parfois intéressant de donner g^* en fonction de $\mathcal{L}_{10}$ le logarithme de vraisemblance de P_1 par rapport à P_0 :

$$g^*(x) = \begin{cases} 1 & \text{si } \mathcal{L}_{10}(x) \geq 0 \\ 0 & \text{sinon} \end{cases}$$

Remarque 5.2. *La fonction $\mathcal{L}_{10}(x)$, dans l'équation précédente, n'est pas du tout la fonction de régression $\eta(x)$ donnée par (5.3). Comme nous le verrons à la Partie II, dans le cadre gaussien, l'utilisation de $\mathcal{L}_{10}(x)$ comme fonction frontière permet de tirer un meilleur parti de la nature géométrique du problème.*

Dans le cas où $K > 2$, si $Y \rightsquigarrow \mathcal{U}(\{1, \dots, K\})$, et $P_1, \dots, P_K$ sont toutes deux à deux équivalentes, la règle optimale associe à x le label $k \in \{1, \dots, K\}$ si x appartient à

$$V_k = \{x \in \mathcal{X} \quad tq \quad \forall j \in \{1, \dots, K\} \quad \mathcal{L}_{kj}(x) \geq 0\}, \tag{5.5}$$

où $\mathcal{L}_{kj}$ est le logarithme rapport de vraisemblance de P_k par rapport à P_j.

Excès de risque. On définit l'excès de risque associé à un classificateur g par :

$$\mathcal{C}(g) - \mathcal{C}(g^*). \tag{5.6}$$

Le théorème classique suivant donne une expression très utile de l'excès de risque.

Theoreme 5.1. *Pour tout classificateur $g : \mathcal{X} \to \{0, 1\}$,*

$$\mathcal{C}(g) - \mathcal{C}(g^*) = \mathbb{E}\left[|2\eta(X) - 1| 1_{g^*(X) \neq g(X)}\right].$$

Démonstration. Notons tout d'abord que $y = g(x)$ si et seulement si ($(y = 1$ et $g(x) = 1)$ ou $(y = 0$ et $g(x) = 0)$). Ainsi,

$$\mathbb{E}[1_{Y=g(X)}|X] = 1_{g(X)=1}\mathbb{E}[1_{Y=1}|X] + 1_{g(X)=0}(1 - \mathbb{E}[1_{Y=1}|X]),$$

et puisque $\mathbb{E}[1_{Y=1}|X] = \eta(X)$, on a :

$$\mathbb{E}[1_{g(X)\neq Y}|X] - \mathbb{E}[1_{g^*(X)\neq Y}|X] = \eta(X)(1_{g^*(X)=1} - 1_{g(X)=1}) + (1-\eta(X))(1_{g^*(X)=0} - 1_{g(X)=0})$$

$$\mathbb{E}[1_{g(X)\neq Y}|X] - \mathbb{E}[1_{g^*(X)\neq Y}|X] = |2\eta(X) - 1|1_{g^*(X)\neq g(X)}. \tag{5.7}$$

Ceci permet de conclure. □

Erreur d'apprentissage. Nous définissons et utilisons une autre quantité permettant de différentier deux règles de classification : l'erreur d'apprentissage.

Définition 5.1. *L'erreur d'apprentissage associée à une règle de classification g est définie par*

$$\mathcal{R}(g) = P(g(X) \neq Y \text{ et } g^*(X) = Y). \tag{5.8}$$

Autrement dit, l'erreur d'apprentissage est la probabilité d'effectuer une erreur de classification avec la règle g alors que la règle g^* n'en fait pas.

Si $K = 2$, $\mathcal{Y} = \{0,1\}$, et $Y \rightsquigarrow \mathcal{U}(\{0,1\})$ alors

$$\mathcal{R}(g) = \frac{1}{2}\left(P_1(g(X) \neq 1 \text{ et } g^*(X) = 1) + P_0(g(X) \neq 0 \text{ et } g^*(X) = 0)\right)$$

On voit assez facilement que pour tout classificateur g,

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq \mathcal{R}(g), \tag{5.9}$$

en effet, on a :

$$\mathcal{C}(g) - \mathcal{C}(g^*) = P(g(X) \neq Y \text{ et } g^*(X) = Y) - P(g(X) = Y \text{ et } g^*(X) \neq Y).$$

Dans le cas gaussien, nous avons obtenu le théorème suivant qui borne inférieurement l'excès de risque par une puissance de l'erreur d'apprentissage.

Theoreme 5.2. *Soient $Y \rightsquigarrow \mathcal{U}(\{0,1\})$, P_0 et P_1 deux mesures gaussiennes et X une variable aléatoire sur $\mathcal{X} = \mathbb{R}^p$ telle que $X|Y \rightsquigarrow YP_1 + (1-Y)P_0$. Soient g^* la règle de Bayes définie par 5.2, $\mathcal{R}$ l'erreur d'apprentissage définie par (5.8) et $\mathcal{C}$ l'erreur de classification donnée par (5.1).*

1. *Si P_0 et P_1 ont la même covariance C et pour moyennes μ_1 et μ_0, alors, pour toute fonction mesurable $g : \mathbb{R}^p \to \{0,1\}$, on a :*

$$\mathcal{C}(g) - \mathcal{C}(g^*) \geq \min\left\{\frac{\sqrt{2\pi}}{2 * 16^2}\|C^{-1/2}m_{10}\|_{\mathbb{R}^p} e^{\frac{\|C^{-1/2}m_{10}\|^2_{\mathbb{R}^p}}{8}}\mathcal{R}(g)^2, \frac{\mathcal{R}(g)}{8}\right\}, \text{ où } m_{10} = \mu_1 - \mu_0.$$

2. *Soit $c_1 > 0$ et $\mathcal{P}(c_1)$ l'ensemble des couples (P,Q) de mesures gaussiennes sur $\mathbb{R}^p$ telles que $d_1(P,Q) > c_1$. Si $(P_1,P_0) \in \mathcal{P}(c_1)$ alors il existe une constante $c(c_1) > 0$ (ne dépendant que de c_1) telle que*

$$\mathcal{C}(g) - \mathcal{C}(g^*) \geq \min\left\{c(c_1)\mathcal{R}(g)^8, \frac{\mathcal{R}(g)}{8}\right\}.$$

Commentaires. Avant de donner la preuve de ce résultat, nous allons le commenter rapidement. Notons que nous avons

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq \frac{1}{2} d_1(P_1, P_0).$$

Aussi, dans le cas où $d_1(P_1, P_0)$ tend vers 0, l'excès de risque ne mesure pas la différence entre la règle g et la règle g^*, mais la proximité des lois P_1 et P_0. L'erreur d'apprentissage n'est pas sensible à ce « facteur d'échelle » . L'exemple suivant en témoigne

Exemple 5.1. *Soient $\mu \geq 0$, et $P_1 = \mathcal{N}(\mu, 1)$ et $P_0 = \mathcal{N}(-\mu, 1)$. Dans ce cas, pour tout $a \in \mathbb{R}$*

$$\mathcal{R}(\mathbb{1}_{[a,\infty[}) = \frac{1}{2} \left(P(0 < \xi + \mu < a) + P(a < \xi - \mu < 0) \right),$$

où $\xi \rightsquigarrow \mathcal{N}(0,1)$; et $d_1(P_1, P_0) \to 0$ si et seulement si $\mu \to 0$ au quel cas

$$\mathcal{R}(\mathbb{1}_{[a,\infty,[}) \to \frac{1}{2} P(\xi \in [0, |a|]).$$

Sous ces conditions l'erreur d'apprentissage de $\mathbb{1}_{[a,\infty,[}$ ne tend vers 0 que si a tend vers 0. En d'autres termes, lorsque $\mu \to 0$, l'erreur d'apprentissage différentie les règles $\mathbb{1}_{[100,\infty,[}$ et $g^* = \mathbb{1}_{[0,\infty,[}$:

$$\inf_{\mu < 50} \mathcal{R}(\mathbb{1}_{[100,\infty[}) \geq \frac{1}{2} P(\xi \in [0, |50|]) \approx \frac{1}{4}$$

alors que

$$\mathcal{C}(\mathbb{1}_{[100,\infty[}) - \mathcal{C}(g^*) \leq \frac{1}{2} d_1(P_1, P_0) \leq \frac{\mu}{\sqrt{2\pi}}.$$

Remarque 5.3. *Par définition, l'excès de risque mesure exactement ce que l'on veut réduire : la probabilité de se tromper. Il n'a donc pas de « défaut » . Le seul problème est qu'il est capable de donner du crédit à n'importe quelle procédure si $d_1(P_1, P_0)$ est suffisamment petit. Ainsi, on ne peut que difficilement mettre en avant le caractère uniformément (sur tout compact) inconsistant d'une méthode donnée. Dans l'exemple précédent, la procédure $g(x) = \mathbb{1}_{[100,\infty[}(x)$ est uniformément (sur $|\mu| \leq 50$) inconsistante selon l'erreur d'apprentissage, mais pas selon l'excès de risque.*

Le théorème précédent nous indique la pertinence de l'erreur d'apprentissage dans le cas où $d_1(P_1, P_0)$ ne tend pas vers 0. D'après l'équation (5.9), si $(g_n)_{n \geq 0}$ est une suite de classificateurs tels que $\mathcal{R}(g_n)$ tend vers zéro, alors $\mathcal{C}(g_n) - \mathcal{C}(g^*)$ tend vers zéro. D'après le théorème 5.2, la réciproque est vraie dans le cas où $K = 2$, $Y \rightsquigarrow \mathcal{U}(\{0,1\})$, si l'on suppose que $(P_1, P_0) \in \mathcal{P}(c)$. Bien entendu, les vitesses de convergence vers zéro de $\mathcal{C}(g_n) - \mathcal{C}(g^*)$ et de $\mathcal{R}(g_n)$ peuvent différer, nous savons seulement que l'une est minorée par l'autre (cf équation (5.9)).

Puisque P_0 et P_1 sont deux mesures gaussiennes, le cas où $A_2(P_1, P_0) \to 0$ correspond au cas où les mesures P_1 et P_0 tendent à être orthogonales.

Démonstration. Notons

$$K_1 = \{x \in \mathbb{R}^p \ : \ g(x) \neq 1 \text{ et } g^*(x) = 1\} \ \text{ et } \ K_0 = \{x \in \mathbb{R}^p \ : \ g(x) \neq 0 \ \text{ et } g^*(x) = 0\}.$$

Ainsi, $\mathcal{R}(g) = \frac{1}{2}(P_1(K_1) + P_0(K_0))$ et au moins une des deux inégalités suivante est vérifiée (c'est le principe du nid de pigeon (pigeonhole principle) :

$$P_1(K_1) \geq \mathcal{R}(g), \quad P_0(K_0) \geq \mathcal{R}(g).$$

Sans restriction de généralité, nous supposerons que $P_1(K_1) \geq \mathcal{R}(g)$. Ceci implique que $P_1(K_1) + P_0(K_1) \geq \mathcal{R}(g)$. Par ailleurs, nous avons

$$\begin{aligned}
\mathcal{C}(g) - \mathcal{C}(g^*) &= P(g \neq Y) - P(g^* \neq Y) \\
&= \frac{1}{2}(P_1(K_1) - P_1(K_0)) + \frac{1}{2}(P_0(K_0) - P_0(K_1)) \\
&\text{(en conditionnant par rapport à } Y) \\
&= \frac{1}{2}((P_1 - P_0)(K_1) + (P_0 - P_1)(K_0)),
\end{aligned}$$

et donc, puisque $g^*(X) = 1$ si et seulement si $dP_1 \geq dP_0$ (par définition de g^* et du fait que $Y \rightsquigarrow \mathcal{U}(\{0,1\})$), nous avons

$$\mathcal{C}(g) - \mathcal{C}(g^*) = \frac{1}{2}\int 1_{K_1 \cup K_0} |dP_1 - dP_0| \geq \frac{1}{2}\int 1_{K_1} |dP_1 - dP_0|. \tag{5.10}$$

Nous allons très largement utiliser les notations et les calculs effectués dans le Chapitre 1 pour obtenir la distance d_1 entre deux mesures gaussiennes. Nous ne les rappelons pas ici, mais nous rappelons que (cf Proposition 1.3 Chapitre 1 Partie I)

$$d_1(P_1, P_0) = 2\mathbb{E}_P\left[e^{f_{10}(P,X)} |\sinh\left(\frac{1}{2}\mathcal{L}_{10}(X)\right)|\right],$$

où P est une mesure de probabilité qui domine P_1 et P_0, $f_{10}(P,X) = \frac{1}{2}\log(\frac{dP_1}{dP}\frac{dP_0}{dP})$ et $\mathcal{L}_{10}(x) = \log(\frac{dP_1}{dP_0}(x))$. Notons par ailleurs que si $K \subset \{x \in \mathbb{R}^p : \mathcal{L}_{10}(x) \geq 0\}$ alors, par la proposition 1.3, on a :

$$P_1(K) - P_0(K) = 2\mathbb{E}_P[1_K e^{f_{10}(P,X)} \sinh(\mathcal{L}_{10}(X)/2)],$$

et donc (5.10) se réécrit

$$\mathcal{C}(g) - \mathcal{C}(g^*) \geq \mathbb{E}[1_{K_1}(X) e^{f_{10}(P,X)} \sinh(\mathcal{L}_{10}(X)/2)]. \tag{5.11}$$

Par la proposition 1.3, on a aussi

$$P_1(K) + P_0(K) = 2\mathbb{E}_P[1_K e^{f_{10}(P,X)} \cosh(\mathcal{L}_{10}(X)/2)],$$

et donc l'inégalité $P_1(K_1) + P_0(K_1) \geq \mathcal{R}(g)$ se réécrit

$$2\mathbb{E}_P[1_{K_1}(X) e^{f_{10}(P,X)} \cosh(\mathcal{L}_{10}(X)/2)] \geq \mathcal{R}(g). \tag{5.12}$$

D'autre part, puisque $d_1(P_1, P_0) \geq c_1$, nous avons :

$$2\mathbb{E}_P[e^{f_{10}(P,X)} |\sinh(\mathcal{L}_{10}(X)/2)|] \geq c_1. \tag{5.13}$$

Dans le reste de la démonstration, il s'agit de combiner (5.12) et (5.13) pour minorer le membre de droite de (5.11). Il s'agit de noter que le membre de gauche de (5.12) et le membre

de droite de (5.11) ne différent que par un facteur 2 et le passage d'un sinh à un cosh, et que ces deux fonctions ne diffèrent fondamentalement (pour notre problème) qu'en zéro. Nous allons décomposer K_1 en deux parties disjointes, ainsi, nous noterons

$$K_1^+ = \{x \in K_1 \ : \ \mathcal{L}_{10}(x) \geq 2\} \ \text{ et } \ K_1^- = \{x \in K_1 \ : \ \mathcal{L}_{10}(x) \leq 2\}.$$

Définissons A et B par :

$$\int_{K_1} e^{f_{10}(P,x)} \sinh(\mathcal{L}_{10}(x)/2) P(dx) = \underbrace{\int_{K_1^+} e^{f_{10}(P,x)} \sinh(\mathcal{L}_{10}(x)/2) P(dx)}_{A} + \underbrace{\int_{K_1^-} e^{f_{10}(P,x)} \sinh(\mathcal{L}_{10}(x)/2) P(dx)}_{B}.$$

D'après (5.12), (et le principe du nid de pigeon) deux cas sont envisageables : soit

$$\mathbb{E}_P[1_{K_1^+}(X) e^{f_{10}(P,x)} \cosh(\mathcal{L}_{10}(X)/2)] \geq \mathcal{R}(g)/4,$$

soit

$$\mathbb{E}_P[1_{K_1^-}(X) e^{f_{10}(P,x)} \cosh(\mathcal{L}_{10}(X)/2)] \geq \mathcal{R}(g)/4. \tag{5.14}$$

Dans le premier cas, puisque $X \in K_1^+$ implique

$$\sinh(\mathcal{L}_{10}(X)/2) \geq \frac{1}{2} \cosh(\mathcal{L}_{10}(X)/2) \ \ (\ln(6) \leq 2),$$

on a $A \geq \mathcal{R}(g)/8$ et alors le théorème (il suffit de noter que $\mathcal{L}_{10}(x) \geq 0$ si $x \in K_1$ et donc $B \geq 0$).

Nous supposerons donc par la suite que c'est l'inégalité (5.14) qui est vérifiée. Ainsi, puisque $\cosh(x) \leq 2$ pour $|x| \leq 1$, nous avons

$$\int_{K_1^-} e^{f_{10}(P,x)} P(dx) \geq \mathcal{R}(g)/8.$$

Ainsi, en posant

$$d\nu = \frac{e^{f_{10}(P,x)} dP}{\int e^{f_{10}(P,x)} dP},$$

ν est une mesure de probabilité sur $\mathbb{R}^p$ et

$$\nu(K_1^-) \geq \mathcal{R}(g)/8. \tag{5.15}$$

Par ailleurs, on a (voir par la définition de f_{10})

$$\int e^{f_{10}(P,x)} dP = \int \sqrt{dP_1 dP_0} = A_2(P_1, P_0)$$

($A_2(P_1, P_0)$ est l'affinité de Hellinger entre P_1 et P_0 définie au Chapitre 1 Partie I) et par conséquent

$$B = A_2(P_1, P_0) \int_0^\infty \nu\left(X \in K_1^- \ \text{ et } |\sinh(\mathcal{L}_{10}(X)/2)| \geq t\right) dt. \tag{5.16}$$

On a :

$$\nu(X \in K_1^-) = \nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}/2)| \leq t\right) + \nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}/2)| \geq t\right).$$

Pour tout $t > 0$, on a :

$$\nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}/2)| \geq t\right) = \nu(X \in K_1^-) - \nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}/2)| \leq t\right)$$

On déduit alors de cette inégalité et de (5.16) que pour tout $\epsilon \geq 0$,

$$\begin{aligned} B &\geq A_2(P_1, P_0) \int_0^\epsilon \nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}(X)/2)| \geq t\right) dt \\ &\geq \epsilon\nu(X \in K_1^-) - A_2(P_1, P_0) \int_0^\epsilon \nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}/2)| \leq t\right) dt) \\ &\geq \epsilon\mathcal{R}(g)/8 - \int_0^\epsilon \nu\left(X \in K_1^- \text{ et } |\sinh(\mathcal{L}_{10}/2)| \leq t\right) dt A_2(P_1, P_0) \end{aligned}$$

où cette dernière inégalité résulte de (5.15). Le reste de la démonstration repose sur le lemme suivant.

Lemme 5.1. *1. L'application w qui à t associe $\sup_{P_1,P_0} \nu(|\sinh(\mathcal{L}_{10}(X)/2)| \leq t)$ vérifie*

$$w(t) \leq \frac{c(c_1)}{A_2(P_1, P_0)} t^{1/7}$$

($c(c_1)$ est une constante positive ne dépendant que de c_1).

2. Dans le cas où $C_1 = C_0 = C$, on a

$$\nu(|\sinh(\mathcal{L}_{10}(X)/2)| \leq t) \leq \frac{4t}{\sqrt{2\pi}\|C^{-1/2} m_{10}\|_{\mathbb{R}^p}}.$$

Nous démontrons ce lemme à la fin de cette preuve, notons que c'est l'équation (5.13) qui joue ici un rôle fondamental.

Dans le cas où $C_1 \neq C_2$,

$$\int_0^\epsilon w(t) dt A_2(P_1, P_0) \leq \tilde{c}(c_1)\epsilon^{1+1/7},$$

et le choix $\epsilon = \left(\frac{\mathcal{R}(g)}{16}\tilde{c}(c_1)\right)^7$ permet de conclure. Dans le cas où $C_1 = C_2$,

$$\int_0^\epsilon \nu(|\sinh(\mathcal{L}_{10}(X)/2)| \leq t) dt \leq \frac{2\epsilon^2}{\sqrt{2\pi}\|C^{-1/2} m_{10}\|_{\mathbb{R}^p}},$$

et le choix $\epsilon = \sqrt{2\pi}\|C^{-1/2} m_{10}\|_{\mathbb{R}^p} \frac{\mathcal{R}(g)}{32 A_2(P_1,P_0)}$ permet de conclure. En effet, rappelons que (cf proposition 1.3) dans le cas où $C_1 = C_0$, on a

$$A_2(P_1, P_0) = \int e^{f_{10}(P,X)} dP = e^{-\frac{\|C^{-1}(\mu_1 - \mu_0)\|^2_{\mathbb{R}^p}}{8}}.$$

□

Nous allons maintenant démontrer le Lemme (5.1)

Démonstration. Pour le point 2, il suffit de noter que si $P_{1/2}$ est une mesure gaussienne de covariance C et de moyenne s_{10}, si X est une variable aléatoire de loi $P_{1/2}$, alors

$$e^{f_{10}(P_{1/2},X)} = e^{-\frac{\|C^{-1}(\mu_1-\mu_0)\|^2_{\mathbb{R}^p}}{8}} \text{ et en loi } \mathcal{L}_{10}(X) \rightsquigarrow \mathcal{N}(0,\sigma^2),$$

où $\sigma^2 = \|C^{-1}(\mu_1-\mu_0)\|^2_{\mathbb{R}^p}$. Ainsi on a bien

$$\nu(|\sinh(\mathcal{L}_{10}(X)/2)| \le t) = P\left(|\mathcal{N}(0,\sigma^2)| \le 2Argsinh(t)\right) \le \frac{4Argsinh(t)}{\sqrt{2\pi}\sigma} \le \frac{4t}{\sqrt{2\pi}\sigma}.$$

Démontrons maintenant le point 1 du lemme.

$$\begin{aligned}\nu(|\sinh(\mathcal{L}_{10}(X)/2)| \le t) &\le \int 1_{|\sinh(\mathcal{L}_{10}(x)/2)|\le t}\left(\frac{dP_1}{dP_0}\right)^{1/2} dP_0/A_2(P_1,P_0).\\ &\le \frac{P_0^{1/2}(|\mathcal{L}_{10}(X)/2| \le t)}{A_2(P_1,P_0)}\\ &\quad \text{(d'après Cauchy-Schwartz et } Argsh(y) \ge y).\end{aligned}$$

Enfin, on peut conclure d'après le point 2 du Théorème 3.5 du Chapitre 3 de la Partie II dont les hypothèses sont vérifiées puisque :

$$\begin{aligned}c_1 &\le d_1(P_1,P_0)\\ &\le 2\sqrt{K(P_0,P_1)}\\ &\quad \text{(d'après l'inégalité (1.35) Chapitre 1 Partie I),}\\ &\le 2\|\mathcal{L}_{10}\|^{1/2}_{L_2(P_0)}\\ &\quad \text{(d'après Cauchy-Schartz).}\end{aligned}$$

□

5.2 Règle Plug-in

Dans le problème de classification supervisé, les lois $(P_k)_{k=1,\dots,K}$ ne sont pas connues mais on dispose d'un échantillon d'apprentissage. Cet échantillon d'apprentissage est composé pour toute classe $k \in \{1,\dots,K\}$, de n observations $((X_1,Y_1),\dots,(X_n,Y_n))$ issues de la loi P. Dans le cas où Y suit une loi uniforme, pour $k \in \{1,\dots,K\}$, on peut supposer observer $X_1^k,\dots,X_n^k$ indépendantes et identiquement distribuées selon la loi P_k. En opposition au cadre bayesien, ce cas convient parfaitement aux échantillons d'apprentissage construits dans le but de fournir beaucoup de données à la moins fournie des classes.

Supposons que $K = 2$. Une manière de construire le classificateur g en utilisant les données d'apprentissage $(X_i,Y_i)_{i=1,\dots,n}$, est de construire un estimateur $\hat{\eta}$ de la fonction de régression η (définie par (5.3)) et de définir

$$\hat{g}(x) = 1 \text{ si } \hat{\eta}(x) \ge 1/2 \quad \text{et 0 sinon.} \tag{5.17}$$

Une telle règle est dite de type plug-in. Dans une première analyse, l'excès de risque est lié à la distance en norme L_1 entre η et $\hat{\eta}$. C'est ce que décrit le théorème suivant.

Theoreme 5.3. *[Plug-in Classifier] Soit $\hat{\eta}$ un estimateur de η, si $\hat{g}$ est donnée par (5.17), on a :*

$$\mathcal{C}(\hat{g}) - \mathcal{C}(g^*) \leq 2\mathbb{E}_X[|\eta(X) - \hat{\eta}(X)|] \quad P^{\otimes n} - ps \tag{5.18}$$

(l'espérance est prise sous la loi de X, et $P^{\otimes n}$ est la loi de l'échantillon d'apprentissage).

Remarque 5.4. *Avant de commencer la démonstration, notons que l'inégalité de ce théorème concerne deux variables aléatoires mesurables par rapport à l'échantillon d'apprentissage : $\mathcal{C}(\hat{f})$ et $\mathbb{E}_X[|\eta(X) - \hat{\eta}(X)|]$. En d'autres termes, cette inégalité ne concerne pas une méthode d'apprentissage de la fonction η mais est intrinsèque à la procédure de classification.*

Démonstration. D'après le Théorème 5.1 :

$$\mathcal{C}(g) - \mathcal{C}(g^*) = \mathbb{E}|2\eta(X) - 1|(1_{g^*(X) \neq g(X)}).$$

De cette égalité, et du fait que si $g(x) \neq g^*(x)$ alors $|\hat{\eta}(x) - \eta(x)| \geq |\eta(x) - 1/2|$, on déduit

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq \int_{x \in \mathcal{X}} 2|\eta(x) - \hat{\eta}(x)| 1_{g(x) \neq g^*(x)} dP_X(x).$$

□

La dernière inégalité de la démonstration nous indique que la différence entre $\eta(x)$ et $\hat{\eta}(x)$ peut être arbitrairement grande tant que $g(x) = g^*(x)$, et que l'erreur faite sur l'estimation de η, quand η est proche de $1/2$, est la plus grave. L'hypothèse de marge permet de contrôler la probabilité que η soit trop proche de $1/2$. Nous dirons que dans le problème de classification supervisée, l'hypothèse de marge d'ordre α est vérifiée s'il existe $\alpha \geq 0$ et $C_0 > 0$ tels que

$$P_X\left(|\eta(X) - 1/2| < t\right) \leq C_0 t^\alpha, \tag{5.19}$$

où P_X est la loi de la nouvelle observation X. Audibert et Tsybakov [6] démontrent que si l'hypothèse de marge d'ordre α est satisfaite et que l'on connaît un estimateur $\hat{\eta}$ de η vérifiant :

$$P^{\otimes n}\left(|\eta(x) - \hat{\eta}(x)| \geq \delta\right) \leq C_1 e^{-C_2 a_p \delta^2} \quad \text{pour } P_X - \text{presque tout } x \tag{5.20}$$

($P^{\otimes n}$ est la vraie loi de l'échantillon d'apprentissage $(X_1, Y_1, \ldots, X_n, Y_n)$ et P_X est la loi de X), alors l'excés de risque converge vers 0 avec une vitesse que l'on peut déterminer.

Theorem 5.4 (Audibert, Tsybakov [6])**.** *Sous les hypothèses définies par les équations (5.20) et (5.19), l'excés de risque converge vers 0 à la vitesse $a_n^{\frac{1+\alpha}{2}}$. Autrement dit, il existe $C > 0$ tel que :*

$$\mathbb{E}_n[\mathcal{C}^b(\hat{g})] - \mathcal{C}^b(g^b) \leq C a_n^{\frac{1+\alpha}{2}}, \tag{5.21}$$

l'espérance $\mathbb{E}_n$ étant ici prise sous la loi de l'échantillon d'apprentissage.

Remarque 5.5. *La borne donnée par l'équation précédente concerne l'excès de risque. Elle n'est pas intrinsèque à la procédure de classification seulement.*

Deuxième partie

Perturbation de règle de décision et classification de courbes

Le travail exposé dans cette partie est motivé en particulier par le problème de segmentation supervisée d'images hyper-spectrales. Ce problème fera l'objet du Chapitre 2 de la Partie III. Il s'agit de trouver une méthode pour identifier au pixel i le label Y_i associé à l'observation X_i. Dans une image hyper-spectrale, X_i appartient à un espace $\mathcal{X}$ de dimension infinie[1]. D'un point de vue plus concret une image hyper-spectrale est une image dans laquelle chaque pixel est un vecteur de grande dimension (on dit aussi que l'image comporte un grand nombre de bandes spectrales).

Pour $k = 1, \ldots, K$, la loi conditionnelle de X Sachant que $Y_i = k$, est P_k. Les lois $(P_k)_{k=1,\ldots,K}$ sont inconnues. Cependant, on dispose d'un échantillon d'apprentissage. Cet échantillon est constitué pour chaque classe k, de la réalisation de n_k variables aléatoires indépendantes de loi P_k. L'objectif est d'utiliser cet échantillon d'apprentissage pour obtenir des informations sur les lois $(P_k)_{k=1,\ldots,K}$ afin de construire une fonction de segmentation de type plug-in. Il ne s'agit donc pas exactement d'apprendre les lois $(P_k)_{k=1,\ldots,K}$ le plus précisément possible mais plutôt d'apprendre ce qui, dans ces lois, permet de construire une fonction de segmentation performante.

Si aucune hypothèse particulière n'est faite sur la similarité de la fréquence d'apparition du label en deux pixels voisins, ce problème est exactement un problème de classification supervisée (cf Introduction de la thèse). Dans le cas contraire, on a intérêt à tirer partie de cette régularité. Quoi qu'il en soit, l'erreur de classification faite en tenant compte de la régularité -par exemple en utilisant l'algorithme multi-échelle de Kolaczyk [46], présenté dans la dernière partie de ce mémoire- est toujours une fonction de l'erreur du problème de classification supervisée.

Le travail de cette partie est donc motivé par la segmentation supervisée d'images hyper-spectrales, mais aussi par la classification de courbes. D'une manière générale cette partie donne une étude du risque de classification en grande dimension dans un cadre gaussien et deux procédures de réduction de dimension résultant de cette étude. La première concerne la classification de courbes et la deuxième la segmentation d'images hyper-spectrales.

Nous nous intéressons plus particulièrement à deux méthodes classiques de classification supervisée : la procédure LDA (Linear Discriminant Analysis) et la procédure QDA (Quadratic Discriminant Analysis). Ces deux méthodes sont basées sur la modélisation de la distribution des individus d'une classe donnée par une loi gaussienne. Dans le cas de la classification binaire, le rapport de vraisemblance entre P_1 et P_0 permet alors de construire la règle optimale liée à la modélisation. Comme nous l'avons vu (Partie I Chapitre 1), cette règle optimale est construite à partir de la distance de Mahalanobis D_i $(i = 1, 0)$. Elle associe a un nouvel individu la classe $k \in \{1, 0\}$ pour laquelle $D_k(x)$ est le plus petit. Bien évidement cette distance est construite à partir de lois qui ne sont pas connues et il faut donc l'estimer. La procédure de classification qui résulte de cet apprentissage est la procédure LDA si les covariances des deux lois modélisant le problème sont les mêmes et la procédure QDA lorsqu'elles sont différentes. La séparation est linéaire dans un cas et quadratique dans l'autre. La difficulté de notre problème réside dans le fait que nous voulons donner une procédure de classification qui soit efficace dans le cas où le nombre n d'observations de l'échantillon d'apprentissage n'est pas nécessairement plus grand que la dimension p du problème.

La grande dimension des données à classifier est donc au coeur de notre problème. Il s'agit de comprendre que la grande dimension du problème modifie fondamentalement la notion d'in-

[1] Nous parlons ici de dimension infinie, mais notre travail s'applique bien évidement aussi au cas de la dimension finie.

formation. D'une part, il est connu depuis les travaux de Rao et Varadarajan [62] que la règle de classification associée à la distance de Mahalanobis peut avoir une erreur de classification tendant vers 0 lorsque la dimension p du problème tend vers l'infini. On peut penser qu'une telle séparation des données facilite la phase d'apprentissage. D'autre part, l'estimation des distances $(D_k)_k$ implique l'estimation d'un nombre de paramètres qui croit avec la dimension p. Ainsi cette estimation est d'autant plus difficile que la dimension p est grande. Il s'agit donc de mesurer l'intérêt d'une dimension supplémentaire en mettant face à face l'information qu'elle apporte et le bruit qu'elle induit. Nous allons quantifier ces phénomènes afin de proposer une procédure adaptée d'estimation de la règle optimale.

Dans cette optique, nous allons mesurer par l'erreur d'apprentissage (voir Chapitre 5 Partie I), l'effet d'une perturbation de la règle de classification optimale. Puisqu'il s'agit d'étudier la différence entre une règle optimale et la règle obtenue lorsque l'on cherche à imiter la règle optimale, notre approche peut être intégrée à la théorie de l'apprentissage et plus particulièrement à l'étude des règles de classification de type plug-in (voir aussi Chapitre 5 Partie I). D'un point de vue théorique notre étude fait suite aux travaux de Bickel et Levina [12]. Nous donnons le lien direct entre l'erreur d'estimation des paramètres et l'erreur d'apprentissage. Dans le cadre des procédures LDA et QDA nous bornons l'erreur d'apprentissage par un terme mesurant l'erreur d'estimation des paramètres. Nous donnons aussi des résultats dans un cadre infini-dimensionnel. D'un point de vue pratique, nous donnons, dans l'esprit de Fan et Fan [32], une technique de réduction de dimension pour la classification basée sur une procédure de tests multiples avec un contrôle du FDR. Nous justifions cette approche par des résultats théoriques et des expérimentations numériques.

Le plan de cette partie est le suivant : dans le premier Chapitre, les théorèmes de perturbation sont motivés, exposés et commentés. Le Chapitre 2 est dédié à l'exposé des deux méthodes pratiques d'estimation de la règle optimale. Celles-ci sont basées sur deux techniques de réduction de dimension. Les méthodes de classification résultantes sont guidées par les principes théoriques du Chapitre 1. Dans le Chapitre 3 nous donnons les démonstrations de nos résultats.

Chapitre 1

Perturbations de règles de décision

> The difference between the normal, rectangular and exponential laws is of course, very great, but the question of what may be termed the stability in form of best critical regions for smaller changes in the frequency law, $p(x_1, \ldots, x_n)$, is of considerable practical importance.
>
> Neymann et Pearson.

Dans ce chapitre nous présentons tous les résultats que nous avons obtenus sur l'erreur d'apprentissage dans le cas des procédures LDA et QDA. Nous commentons ces résultats. Nous définissons et étudions d'abord le cas fini dimensionnel puis le cas des espaces de Banach infinis dimensionnels. Dans le cadre de la procédure LDA, Nous donnons une condition nécessaire et une condition suffisante pour que l'erreur d'apprentissage tende vers 0 *lorsque les lois de probabilité des deux groupes tendent à être orthogonales.*

1.1 Introduction

Dans le cas de deux classes, la classification supervisée de données gaussiennes est le problème suivant. Nous supposons avoir un échantillon d'apprentissage composé de n variables aléatoires indépendantes ayant pour loi soit $P_0 = \gamma_{C_0,\mu_0}$, soit $P_1 = \gamma_{C_1,\mu_1}$ où C_0 et C_1 sont des matrices symétriques définies positives et $\gamma_{C,\mu}$ est la mesure gaussienne sur $\mathcal{X} = \mathbb{R}^p$ de moyenne μ et de covariance C. Par exemple, dans le cadre de l'application médicale, on dispose d'une série de spectres. Chacun des spectres est résumé par p bandes spectrales, et est issu soit d'une tumeur 1 soit d'une tumeur 0. Nous voulons alors construire une procédure permettant de décider si un nouveau vecteur $X \in \mathbb{R}^p$ est issu de P_0 ou de P_1. Du point de vue de l'application médicale cette procédure doit permettre d'associer à un nouveau spectre, un type de tumeur donné.

Les lois P_0 et P_1 ne sont parfaitement connues que si l'échantillon d'apprentissage est infini.

Dans ce cas idéal, il s'agit, au vu de l'observation de X un vecteur aléatoire (un élément de $\mathcal{X} = \mathbb{R}^p$) de tester les hypothèses

$$H_0 \ : \ X \rightsquigarrow P_0 \text{ contre } H_1 \ : \ X \rightsquigarrow P_1, \tag{1.1}$$

tout en contrôlant la somme des erreurs de seconde et de première espèce. La procédure optimale, consiste à rejeter H_0 si X appartient à

$$V = \{x \in \mathbb{R}^p \text{ tq } \mathcal{L}_{10}(x) \geq 0\}, \tag{1.2}$$

où $\mathcal{L}_{10}(x)$ est le logarithme du rapport de vraisemblance entre P_1 et P_0 (le logarithme de la dérivée de Radon-Nikodym de P_1 par rapport à P_0). En d'autres termes, la partie V définie par (1.2) minimise

$$\frac{1}{2}\left(P_0(X \in V) + P_1(X \notin V)\right). \tag{1.3}$$

Dans le cas de plusieurs classes, on observe X suivant une loi $P \in \{P_1, \ldots, P_K\}$. Il existe une partition de l'espace considéré $\mathcal{X} = \mathbb{R}^p$, en $(V_k)_{k=1,\ldots,K}$ qui permet de décrire la règle de décision optimale : affecter la nouvelle observation X à la classe k si $X \in V_k$. Cette partition peut être définie par K^2 fonctions $\mathcal{L}_{k_1,k_2}$ telles que $\mathcal{L}_{k_1,k_2} = -\mathcal{L}_{k_2,k_1}$ et $\mathcal{L}_{k,k} = 0$:

$$V_k = \{x \in \mathcal{X} \quad tq \ \ \forall j \in \{1, \ldots, K\} \ \ \mathcal{L}_{kj}(x) \geq 0\}, \tag{1.4}$$

où $\mathcal{L}_{kj}$ est le logarithme du rapport de vraisemblance entre P_k et P_j. En effet, si P est une mesure qui domine $P_1, \ldots, P_K$, on a :

$$\mathcal{L}_{k,i}(x) = \log\left(\frac{dP_k}{dP}(x)\right) - \log\left(\frac{dP_i}{dP}(x)\right),$$

et

$$V_k = \left\{x \in \mathcal{X} \quad tq \ \ k = \text{Argmax}_i \frac{dP_i}{dP}(x)\right\} = \left\{x \in \mathcal{X} \quad tq \ \ \forall i \in \{1, \ldots, K\} \ \ \frac{dP_k}{dP}(x) \geq \frac{dP_i}{dP}(x)\right\},$$

ce qui implique (1.4).
Par la suite, dans ce chapitre, sauf mention du contraire, nous nous restreindrons à l'étude du cas $K = 2$.

Comme nous allons le voir, les procédures LDA et QDA sont basées sur l'estimation de $\mathcal{L}_{10}$ à partir de l'échantillon d'apprentissage. Nous souhaitons consacrer un chapitre à l'étude théorique des propriétés d'une règle construite avec un estimateur $\hat{\mathcal{L}}_{10}$ de $\mathcal{L}_{10}$. Nous avons voulu donner des idées théoriques précises sur ce qui est à l'origine d'un « bon apprentissage » . Nous souhaitons établir le parallèle entre le **problème d'estimation** d'un paramètre de $\mathbb{R}^p$ avec une erreur mesurée en norme l^2 (Cf Annexe A), et le problème d'estimation de $\mathcal{L}_{10}$ pour la classification avec une erreur mesurée en terme d'erreur d'apprentissage. Ce deuxième type de problème sera appelé le **problème d'apprentissage**. Rappelons que l'erreur d'apprentissage a été définie et étudiée au Chapitre 5 de la Partie I (Définition 5.1). Nous supposerons que les classes 0 et 1 ont la même probabilité d'apparition a priori (cela est motivé dans le Chapitre 5 Partie I). Ainsi, dans toute la suite, si $g : \mathcal{X} \to \{0, 1\}$, nous noterons

$$\mathcal{R}(g) = \frac{1}{2}\left(P_1(g(X) \neq 1 \text{ et } g^*(X) = 1) + P_0(g(X) \neq 0 \text{ et } g^*(X) = 0)\right), \tag{1.5}$$

où g^* est la règle de Bayes (définie Chapitre 5 Partie I). Rappelons que (cf Chapitre 5 de la Partie I) cette quantité est un majorant de l'excès de risque et que dans le cas gaussien l'excès de risque est borné inférieurement par un polynôme nul en zéro de l'erreur d'apprentissage (cf Theoreme 5.2 Chapitre 5 Partie I).

Dans toute cette partie, si $F \in \mathbb{R}^p$ et γ est une mesure gaussienne, nous noterons $\|F\|_{L_2(\gamma)}$ la norme L_2 de l'application $x \in \mathbb{R}^p \to \langle F, x\rangle_{\mathbb{R}^p}$. Notons que si γ est la mesure gaussienne centrée de covariance C, alors $\|F\|_{L_2(\gamma)} = \|C^{1/2}F\|_{\mathbb{R}^p}$.

Dans un premier temps (Section 2), nous explicitons les conséquences d'une erreur d'estimation des paramètres sur l'erreur d'apprentissage de la procédure de classification plug-in correspondante. Dans le Théorème 1.1 nous bornons supérieurement l'erreur d'apprentissage par un terme mesurant directement l'erreur d'estimation des paramètres. Dans un deuxième temps (Section 3), nous donnons des résultats du même type dans le cadre plus élaboré de la procédure QDA. Afin de donner un cadre asymptotique rigoureux à notre étude, nous formulons dans un troisième temps nos théorèmes dans un cadre infini-dimensionnel. C'est l'objet de la Section 4. Nous exposons dans la Section 5 quelques conjectures et perspectives.

1.2 Régle linéaire, perturbation linéaire : LDA.

Nous allons donner et commenter dans les trois sous-sections qui suivent, trois résultats que nous avons obtenu concernant la procédure LDA.

1.2.1 Définition de la procédure LDA et résultat principal

Nous allons supposer que C est une matrice symétrique définie-positive. Dans le cas où $C_1 = C_0 = C$, $\mathcal{L}_{10}(x) = \mathcal{L}_{10}^A(x)$ est une application affine sur $\mathbb{R}^p$:

$$\mathcal{L}_{10}^A(x) = \langle F_{10}, x - s_{10}\rangle_{\mathbb{R}^p} \text{ où } s_{10} = \frac{\mu_1 + \mu_0}{2},\ F_{10} = C^{-1}m_{10} \text{ et } m_{10} = \mu_1 - \mu_0. \tag{1.6}$$

Dans le problème de classification supervisée, l'échantillon d'apprentissage doit nous permettre de construire des estimateurs $\hat{F}_{10}$ et $\hat{s}_{10}$ de F_{10} et s_{10}. Nous décidons alors que l'observation X est issu de P_1 si elle appartient à

$$\hat{V} = \left\{x \in \mathbb{R}^p \text{ tq } \widehat{\mathcal{L}}_{10}^A(x) \geq 0\right\}, \tag{1.7}$$

où $\widehat{\mathcal{L}}_{10}^A(x)$ est définit en substituant $\hat{F}_{10}$ et $\hat{s}_{10}$ à F_{10} et s_{10} dans (1.6). On peut définir, dans la géométrie de $L_2(\gamma_C)$, l'angle $\alpha \in [0, \pi]$ entre F_{10} et $\hat{F}_{10}$ par

$$\alpha = \arccos\left(\frac{\langle \hat{F}_{10}, F_{10}\rangle_{L_2(\gamma_C)}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}\|F_{10}\|_{L_2(\gamma_C)}}\right). \tag{1.8}$$

Cet angle va jouer un rôle déterminant dans la suite.

Theoreme 1.1. *Soient $\hat{F}_{10}$ et $\hat{s}_{10}$ deux vecteurs de $\mathbb{R}^p$ et $\widehat{\mathcal{L}}_{10}^A(x)$ définie en substituant $\hat{F}_{10}$ et $\hat{s}_{10}$ à F_{10} et s_{10} dans (1.6). Soient P_1 et P_0 deux mesures gaussiennes sur $\mathcal{X} = \mathbb{R}^p$ de même covariance C et de moyennes respectivement μ_1 et μ_0. Alors, si $\hat{V}$ est la partie de $\mathbb{R}^p$ définie par (1.7), on a :*

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_C)}}$$

où

$$\mathcal{E} = \left(\frac{4\|F_{10}\|_{L_2(\gamma_C)}}{\sqrt{\pi}\|\hat{F}_{10}\|_{L_2(\gamma_C)}} |\langle \hat{F}_{10}, \hat{s}_{10} - s_{10} \rangle_{\mathbb{R}^p}| + \|F_{10} - \hat{F}_{10}\|_{L_2(\mathbb{R}^p, \gamma_C)} \right), \tag{1.9}$$

et $\mathcal{R}$ est l'erreur d'apprentissage donnée par l'équation (1.5). De plus, si $|\langle \hat{F}_{10}, \hat{s}_{10} - s_{10} \rangle_{\mathbb{R}^p}| \leq \frac{1}{4}|\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)}|$ et $\alpha \leq \pi/4$ (α est défini par (1.8)), alors

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{32}} \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_C)}}. \tag{1.10}$$

La démonstration de ce théorème est donnée dans le Chapitre 3 à la Sous-section 3.1.3. Il résulte du Théorème 3.1 obtenu par des méthodes géométriques simples. Ce dernier théorème est plus général, mais aussi moins simple à formuler et à appréhender. Son énoncé et sa démonstration, dont tous nos résultats concernant la procédure LDA résultent, sont reportés au Chapitre 3. Nous allons maintenant commenter le Théorème 1.1.

Commentaires généraux. Nous commençons par donner quelques commentaires simples et assez généraux.

Si l'on note

$$\delta = \hat{F}_{10} - F_{10} \text{ et } d_0 = \langle \hat{F}_{10}, s_{10} - \hat{s}_{10} \rangle_{\mathbb{R}^p}, \tag{1.11}$$

on a

$$\hat{\mathcal{L}}_{10}(x) = \mathcal{L}_{10}(x) + \langle \delta, x - s_{10} \rangle_{\mathbb{R}^p} + d_0.$$

Ainsi nous parlerons dans la suite de perturbation linéaire (ou affine) de la règle optimale. Le théorème précédent est un théorème qui résulte de l'étude des perturbations affines d'une règle affine.

La quantité $r = \|F_{10}\|_{L_2(\gamma_C)}$ mesure la séparation « théorique » des données. En effet, comme nous l'avons vu (Chapitre 1 Partie I) c'est la distance L_1 entre P_1 et P_0 qui mesure, au sens des tests, la séparation des données. Par ailleurs, d'après la Proposition 1.3 Chapitre 1 Partie I, nous avons

$$d_1(P_1, P_0) = \Phi\left(-\frac{1}{2}r\right) - \Phi\left(\frac{1}{2}r\right) \sim r \text{ quand } r \to 0,$$

($\Phi(x)$ est la fonction de répartition d'une loi normale centrée réduite) c'est-à-dire lorsque les données tendent à être indistinguables ($d_1(P_1, P_0) \to 0$). Lorsque les données tendent à être parfaitement séparées ($d_1(P_1, P_0) \to 1$) alors $r \to \infty$ et

$$d_1(P_1, P_0) \sim 1 - \frac{2e^{-\frac{r^2}{8}}}{r\sqrt{2\pi}}.$$

Ainsi, si $\mathcal{E}$ mesure l'erreur d'estimation, les termes

$$\frac{1}{\|F_{10}\|_{L_2(\gamma_C)}} \quad \text{et} \quad e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{32}} \tag{1.12}$$

dans les majorations (1.9) et (1.10), sont des termes propres au problème de test des hypothèses définies par (1.1). Lorsque $\|F_{10}\|^2_{L_2(\gamma_C)}$ est grand, les données sont mieux séparées et les termes de (1.12) mesurent l'influence de cette séparation des données sur l'excès de risque. Nous pensons que lorsque $\|F_{10}\|^2_{L_2(\gamma_C)}$ tend vers 0, le terme $\frac{1}{\|F_{10}\|_{L_2(\gamma_C)}}$ est lié à la mesure du risque utilisée (erreur d'apprentissage). Il n'est pas correct de penser que le problème de classification est plus difficile (au sens de l'excès de risque) lorsque les données ne sont pas parfaitement séparées. En effet, on a (cf Chapitre 5 Partie I)

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq \frac{1}{2} d_1(P_1, P_0),$$

et donc, si $d_1(P_0, P_1) \to 0$, alors $\sup_g \mathcal{C}(g) - \mathcal{C}(g^*)$ tend vers zéro. Comme nous l'avons mentionné (cf Théorème 5.2 et Remarque 5.3 Chapitre 5 Partie I) la mesure de risque utilisée (l'erreur d'apprentissage) ne se comporte comme l'excès de risque que si $d_1(P_0, P_1)$ ne tend pas vers 0.

L'échantillon d'apprentissage doit être utilisé pour la construction d'estimateurs $\hat{F}_{10}$ et $\hat{s}_{10}$ de F_{10} et s_{10}. Le théorème précédent permet de quantifier ce que l'intuition indique clairement : une bonne estimation des paramètres F_{10} et s_{10} (ou plus indirectement de μ_1, μ_0 et C) permet d'obtenir une bonne règle de classification. Ces estimateurs doivent avoir comme propriété principale d'offrir une bonne classification et donc un excès de risque petit. En vertu du théorème précédant et des propriétés de l'erreur d'apprentissage (cf Chapitre 5 Partie I) l'espérance de l'excès de risque vérifie

$$\mathbb{E}_{P^{\otimes n}}[\mathcal{C}(\mathbb{1}_{\hat{V}}) - \mathcal{C}(\mathbb{1}_V)] \leq \mathbb{E}_{P^{\otimes n}}[\mathcal{R}(\mathbb{1}_{\hat{V}})] \leq \frac{\mathbb{E}_{P^{\otimes n}}[\mathcal{E}]}{\|F_{10}\|_{L_2(\gamma_C)}}, \tag{1.13}$$

où $P^{\otimes n}$ est la loi de l'échantillon d'apprentissage (rappelons que $\mathcal{C}(g)$ est l'erreur de classification associée à un classificateur g, V est donné par (1.2) et $\mathbb{1}_V$ est donc la règle de Bayes).

Il semblerait que peu de choses soient dites sur la qualité de la procédure LDA (procédure plug-in) par rapport à la règle optimale (règle de bayes). Le résultat classiquement utilisé (voir par exemple Anderson et Bahadur [3]) pour montrer la consistance d'une règle utilisant des estimateurs $\hat{F}_{10} = \widehat{C^{-1}}\hat{m}_{10} = \widehat{C^{-1}}(\hat{\mu}_1 - \hat{\mu}_0)$ et $\hat{s}_{10} = (\hat{\mu}_1 + \hat{\mu}_0)/2$ est que la probabilité que $X \rightsquigarrow \gamma_{C,\mu_0}$ (qui appartient alors à la classe 0) soit affecté à la classe 1 est :

$$P\left(\langle \hat{F}_{10}, C^{1/2}\xi\rangle_{\mathbb{R}^p} \geq \langle \hat{s}_{10} - \mu_0, \hat{F}_{10}\rangle_{\mathbb{R}^p} | \mathcal{A}\right) = 1 - \Phi\left(\frac{\langle \hat{s}_{10} - \mu_0, \hat{F}_{10}\rangle_{\mathbb{R}^p}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}\right), \tag{1.14}$$

où $\mathcal{A}$ est la tribu engendrée par l'échantillon d'apprentissage, ξ est un vecteur gaussien centré réduit dans $\mathbb{R}^p$. Notons que la démonstration de (1.14) est triviale. Nous pensons qu'une analyse directe de ce terme d'erreur ne permet pas d'utiliser pleinement la nature géométrique du problème. Par ailleurs, ce terme doit être comparé à l'erreur commise par la règle optimale (la règle de Bayes) pour mesurer la qualité de l'apprentissage. Notons que pour la procédure LDA en

grande dimension, une analyse du pire des excès de risque a été faite à partir de (1.14) par Bickel et Levina [12] pour un choix particulier de $\hat{F}_{10}$ et $\hat{s}_{10}$. Notre théorème, parce qu'intrinsèque à la procédure de classification, diffère singulièrement du type de résultat qu'ils obtiennent. En outre il va nous permettre d'établir un lien clair entre réduction de dimension en classification et estimation par seuillage.

Partie constante de la perturbation. L'erreur, due à la partie constante de la perturbation (d_0 dans l'équation (1.11)), est mesurée par

$$\frac{4}{\sqrt{\pi}} \left| \langle \frac{\hat{F}_{10}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}, \hat{s}_{10} - s_{10} \rangle_{\mathbb{R}^p} \right|.$$

Afin de donner une première analyse simple de ce terme nous allons supposer que $\hat{F}_{10}$ et $\hat{s}_{10}$ sont indépendants. Cette indépendance peut être obtenue en réservant une partie de l'échantillon d'apprentissage à l'estimation de F_{10} et une partie à l'estimation de s_{10}. Dans ce cas, si n' observations de l'échantillon d'apprentissage ont été utilisées pour la construction de $\hat{s}_{10}$, et si $\hat{s}_{10} = (\bar{\mu}_1 + \bar{\mu}_0)/2$ ($\bar{\mu}_i$ est la moyenne empirique des observations du groupe i), alors, sachant le reste de l'échantillon d'apprentissage, on a :

$$\left\langle \frac{\hat{F}_{10}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}, \hat{s}_{10} - s_{10} \right\rangle_{\mathbb{R}^p} \rightsquigarrow \gamma_{\frac{1}{n'}}.$$

Ceci implique, que

$$\mathbb{E}_{P^{\otimes n}} \left[\frac{4}{\sqrt{\pi}\|\hat{F}_{10}\|_{L_2(\gamma_C)}} |\langle \hat{F}_{10}, \hat{s}_{10} - s_{10} \rangle_{\mathbb{R}^p}| \right] \leq \frac{8}{\sqrt{2n'\pi}}.$$

En définitive la difficulté du problème ne provient absolument pas de la partie constante de la perturbation, mais bien de la partie linéaire. C'est sur cette partie linéaire que nous allons plus longuement discuter par la suite.

Les conditions données pour la deuxième inégalité (1.10) du théorème ne sont pas du tout marginales. La deuxième condition porte sur l'angle α fait entre F_{10} et $\hat{F}_{10}$ dans la géométrie induite par le produit scalaire de $L_2(\gamma_C)$ (voire équation (1.8)) : il s'agit que $\alpha \leq \frac{\pi}{4}$. On peut penser que, avec une grande probabilité, un bon estimateur $\hat{F}_{10}$ de F_{10} vérifie ces hypothèses. La première condition est vérifiée si la deuxième l'est et que l'on a :

$$\left| \left\langle \frac{\hat{F}_{10}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}, s_{10} - \hat{s}_{10} \right\rangle_{\mathbb{R}^p} \right| \leq \frac{\sqrt{2}}{8} \|F_{10}\|_{L_2(\gamma_C)}.$$

Si par exemple $\hat{s}_{10} = (\bar{\mu}_1 + \bar{\mu}_0)/2$ et que l'échantillon d'apprentissage est composé de n' observations utilisées uniquement pour l'estimation de s_{10}, alors sachant le reste de l'échantillon d'apprentissage, $\langle \frac{\hat{F}_{10}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}, s_{10} - \hat{s}_{10} \rangle_{\mathbb{R}^p} \rightsquigarrow \gamma_{\frac{1}{n'}}$ et la condition précédente est réalisée avec une probabilité

$$\frac{1}{2}\Phi\left(\frac{\sqrt{2}}{8} \|F_{10}\|_{L_2(\gamma_C)} n' \right).$$

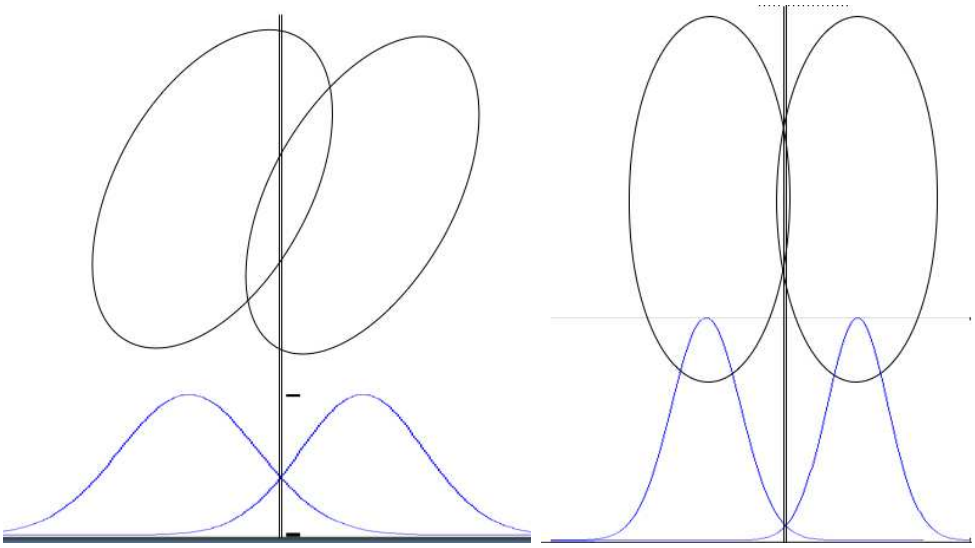


FIG. 1.1 – Séparation des données avant et après image par C

Partie linéaire de la perturbation. Comme nous l'expliquons dans la preuve du théorème, c'est l'angle α définit par (1.8) qui mesure le mieux l'erreur due à la partie linéaire de la perturbation. Ainsi, on peut saisir quelle est l'imperfection dans la majoration du théorème énoncé. En effet si $\beta \in \mathbb{R}$, et $\hat{F}_{10} = \beta F_{10}$, l'erreur d'apprentissage est nulle alors que la borne (1.9) peut être arbitrairement grande. Nous pensons qu'une étude de l'estimation d'une direction (un paramètre de la sphère $\mathbb{S}^{p-1}$) en grande dimension devrait être faite, mais nous voulons ici donner le lien entre l'estimation de F_{10} comme vecteur de $\mathbb{R}^p$ et l'erreur d'apprentissage. Par ailleurs, cette invariance de l'erreur par dilatation n'a lieu que dans la direction F_{10} qui est inconnue et il paraît donc délicat d'en faire usage de manière directe.

Réduction de dimension et estimation par seuillage. Si le théorème précédent peut avoir un intérêt pour des dimensions p petites, nous voulons étudier les cas où p est grand, et éventuellement être en mesure de faire tendre p vers l'infini.

Donnons un exemple simple pour illustrer l'intérêt du lien entre les problèmes d'estimation et d'apprentissage. Soit $\sigma > 0$, supposons que l'on observe X suivant une loi $\gamma_{\frac{1}{\sqrt{n}} I_p, F_{10}}$, que $C = I_p$ et que s_{10} est connu. Dans le problème d'estimation de F_{10} pour la classification, l'erreur $\mathcal{E}$ est $\|F_{10} - \hat{F}_{10}\|_{L_2(\gamma_C)} = \|\hat{F}_{10} - F_{10}\|_{\mathbb{R}^p}$. Ainsi, dans ce cas, le problème est exactement un problème de régression avec une erreur mesurée en norme l^2. Supposons que la dimension p soit assujettie à tendre vers l'infini. Si la décroissances des coefficients de F_{10} est assez forte, par exemple $F_{10} \in l^q(R)$ avec $q < 2$, alors (cf Annexe A), on peut obtenir une bonne méthode d'estimation de F_{10} en fixant à 0 les coefficients de trop faible amplitude. C'est l'estimation par seuillage. Il y a alors concernant l'estimation de F_{10} deux réelles difficultés :

1. Trouver une base orthonormale de $\mathbb{R}^p$ dans laquelle, a priori, la décroissance des coefficients de F_{10} est la plus forte.
2. Choisir un seuil à partir duquel les petites valeurs observées sont remplacées par zéro.

Les deux problèmes concernant l'estimation se traduisent par deux problèmes dans le cadre de l'apprentissage. Le premier problème est un problème d'approximation. En terme de classifica-

tion cette approximation correspond à la recherche d'un sous-espace de dimension donnée dans lequel les données sont bien séparées. Remplacer un coefficient par zéro dans l'estimation de F_{10} correspond exactement à sélectionner un espace de dimension réduite dans lequel la règle de classification agit. Trouver le seuil à partir duquel les petites valeurs observées sont remplacées par zéro revient donc à trouver la dimension de l'espace dans lequel on effectuera finalement la classification. En terme d'estimation, trouver cette dimension limite revient à faire un compromis entre le biais et la variance d'estimation. En terme d'apprentissage ce choix de dimension peut être interprété comme un compromis entre la qualité de séparation des données et la taille des données.

Dans le cas où l'on observe X suivant une loi $\gamma_{C/n,m_{10}}$ (ou X^i, $i = 0, 1$, suivant une loi $\gamma_{C/n,\mu_i}$) et que $C \neq I_p$ est connu, le problème se ramène au cas particulier précédent grâce à la transformation $x \to C^{-1/2}x$. Lorsque C est inconnu, le parallèle avec le problème d'estimation est plus délicat, puisque l'erreur $\mathcal{E}$ dépend de C.

Dans les méthodes de seuillage (cf Annexe A) une quantité importante qu'il s'agit de réduire est le biais d'approximation. Supposons que $\|C^{-1/2}m_{10}\|_{\mathbb{R}^p} \leq 1$. Si l'on cherche notre estimateur $\hat{F}_{10}$ dans un sous-espace L_k de $\mathbb{R}^p$ de dimension $k < p$, alors ce biais d'approximation est dans le pire des cas

$$\sup_{\|C^{-1/2}m_{10}\|_{\mathbb{R}^p}\leq 1} \inf_{F_{10}^k} \|F_{10} - F_{10}^k\|_{L_2(\gamma_C)}.$$

Supposons que C est fixé et connu, que A est une matrice symétrique définie positive et que $\eta_1 \geq \ldots, \geq \eta_p$ sont les valeurs propres ordonnées associées aux vecteurs propres $e_1, \ldots, e_p$ de A. On montre (Théorème de Kolmogorov, voir par exemple Chapitre introductif de [58]) que si $k \leq p$

$$\inf_{L_k} \sup_{\|AC^{-1/2}m_{10}\|_{\mathbb{R}^p}\leq 1} \inf_{F_{10}^k \in L_k} \|F_{10} - F_{10}^k\|_{L_2(\gamma_C)} = \frac{1}{\eta_{k+1}},$$

où l'infimum est prit sur tous les sous-espaces L_k de dimension k de $\mathbb{R}^p$. Cet infimum est atteint pour l'espace engendré par $e_1, \ldots, e_k$. La base de $L_2(\gamma_C)$ correspondante est $(f_i)_{i=1,\ldots,p}$ où $f_i = C^{-1/2}e_i$. Ainsi, la recherche d'une base de $L_2(\gamma_C)$ dans laquelle le biais d'approximation de $F_{10} = C^{-1}(\mu_1 - \mu_0)$ est faible dans le pire des cas peut être effectuée en trouvant la base orthonormale $(e_i)_{i=1,\ldots,p}$ de $\mathbb{R}^p$ qui diagonalise la matrice A puis en prenant l'image par C des éléments de cette base (ceci est illustré par la Figure 1.1). Ceci peut constituer une première étape de recherche.

La deuxième étape de réduction du biais d'approximation peut consister à ordonner les éléments de cette base par valeurs décroissantes de $|\langle C^{-1}(\mu_1 - \mu_0), C^{-1/2}e_i\rangle_{L_2(\gamma_C)}| = |\langle C^{-1/2}(\mu_1 - \mu_0), e_i\rangle_{\mathbb{R}^p}|$.

Notons que ces deux étapes telles qu'elles sont décrites ici sont bien sûr théoriques puisqu'elles font appel à des paramètres inconnus (C, μ_0 et μ_1). La procédure nécessite l'identification de la structure de corrélation entre les attributs. Comme nous allons le voir par la suite (cf Proposition 1.1) cette identification, lorsque $p > n$, ne doit pas être faite de manière directe. La démarche que nous adopterons est la plus élémentaire. Elle consiste à supposer que l'on connaît une base dans laquelle la matrice de covariance C est diagonale. Cette hypothèse est équivalente à supposer que dans cette base les covariables (ou attributs) sont indépendantes et repose au fond sur un

principe d'approximation.

Sélection des directions discriminantes. La deuxième partie de la procédure optimale est souvent négligée. Ainsi certaines procédures de classification couramment utilisées consistent souvent en une simple Analyse en Composantes Principales (diagonalisation de la matrice de covariance empirique) couplée avec une réduction de dimension basée sur la sélection des valeurs propres les plus fortes de la matrice de covariance empirique. Ce type de stratégie est pourtant contre-intuitif. En effet, une telle approche n'utilise pas l'information contenue dans les labels et les plus grandes valeurs propres de cette matrice correspondent à des directions selon lesquelles les données ne sont pas forcement bien séparées.

Nous rappelons que la variabilité d'un nuage de points de $\mathbb{R}^p$ regroupant deux populations issues de lois différentes peut être décomposée en une variabilité intra (moyenne pondérée des matrices de covariance au sein de chaque groupe) qui reflète la variabilité au sein de chaque groupe, et une variabilité inter qui reflète la variabilité entre les groupes. La variabilité totale est la somme de ces variabilités. C'est une quantité empirique. Une grande variabilité inter et une faible variabilité intra contribuent à séparer les données des différents groupes. Notons que ce sont les quantités théoriques associées qui mesurent réellement la séparation des données telles qu'elles sont susceptibles d'apparaître dans la phase de classification.

Il n'est pas raisonnable de chercher une procédure de réduction de dimension qui privilégie des directions dans lesquelles les deux variabilités sont importantes, mais il est du ressort du bon sens de privilégier celles dans lesquelles la variabilité intra est faible et la variabilité inter est forte. Cette heuristique est à l'origine de la méthode de Fisher [34] qui repose sur la maximisation du quotient de Rayleigh (voir par exemple [35]). Le quotient de Rayleigh est le rapport de la variabilité inter sur la variabilité intra. La méthode de Fisher repose sur l'utilisation de la version empirique de ce rapport et ne constitue pas une réduction de dimension. L'intérêt de la version « théorique » de ce rapport est confirmé par le théorème énoncé. En effet, trouver les directions pour lesquelles $|\langle C^{-1/2}(\mu_1 - \mu_0), e_i\rangle_{\mathbb{R}^p}|^2$ est grand revient à trouver les directions pour lesquelles la version « théorique » du rapport entre variabilité inter et variabilité intra est important. Ceci revient à trouver des directions pour lesquelles la version « théorique » du quotient de Rayleigh est importante. Dans le cas où $p = 1$, $(\mu_1 - \mu_0)$ mesure la variabilité inter « théorique » et $C^{-1/2}$ (qui est un réel) la variabilité intra « théorique » . Comme nous le verrons par la suite, si la version théorique du rapport entre la variabilité intra et la variabilité inter est une quantité d'intérêt il n'en est pas forcement de même des quantité empiriques correspondantes si celles si amène à une mauvaise estimation de leurs contreparties théoriques.

Nécessité d'une règle symétrique. La démonstration du Théorème 1.1 donnée à la Section 3 nous donne une dernière indication. Le terme d'erreur d_0, défini par (1.11), est à l'origine d'un déséquilibre de l'importance donnée aux deux types d'erreurs. Plus précisément si $d_0 > 0$ alors l'erreur consistant à affecter la nouvelle donnée X au groupe 0 à tort prend de l'importance au détriment de l'autre type d'erreur, celle consistant à affecter la nouvelle donnée X au groupe 1 à tort.

C'est pour cela que nous proposons de donner une estimation de F_{10} et s_{10} qui soit une fonction symétrique des deux parties de l'échantillon d'apprentissage. Rappelons que dans le cadre de la classification supervisée à deux classes une règle de décision est une fonction de X^0 et X^1 les données d'apprentissages respectivement des groupes 0 et 1, et d'une (nouvelle) observation

X indépendante de l'échantillon d'apprentissage. Notons $f(X^0, X^1, X)$ cette fonction. Dans le cas d'une règle non randomisée, cette fonction est à valeur dans $\{0; 1\}$. Ceci nous amène à la définition suivante.

Définition 1.1 (apprentissage symétrique)**.** *Dans le problème de classification supervisée à deux classes, la règle de décision $f(X^0, X^1, X)$ est dite* ***basée sur un apprentissage symétrique****, si on a en loi l'égalité*

$$f(X^0, X^1, x) = 1 - f(X^1, X^0, x) \quad \textit{pour } P_X \textit{ presque tout } x. \tag{1.15}$$

Cette propriété est vérifiée lorsque l'on utilise un apprentissage qui repose sur les estimateurs empiriques traditionnels (règle de Fisher), mais les méthodes de réduction de dimension que nous allons utiliser peuvent détruire cette symétrie. Si l'apprentissage est symétrique, alors d_0 a une loi symétrique, et donc aucun des deux groupes n'est privilégié par la procédure de classification.

Cette remarque peut aussi être illustrée par l'heuristique suivante concernant le cas de données parfaitement séparées. Nous expliquerons dans la suite en quoi deux mesures gaussiennes qui ne sont pas équivalentes sont orthogonales. Dans la situation où les mesures sont orthogonales, les données sont parfaitement séparées et la règle de décision optimale peut être construite à partir d'ensembles de mesure 1 pour l'une des deux mesures et 0 pour l'autre (cf Chapitre 1 Partie I). Dans le problème de classification, ces ensembles sont inconnus et le fait que les données soient bien séparées[1] l'est aussi. Nous allons donner un résultat permettant de dire si une méthode est adaptée à l'étude de données bien séparées ou si elle ne l'est pas (cf Théorème 1.2). Pour construire une règle qui soit adaptée au cas de données bien séparées, il ne faut privilégier aucun des deux groupes et ne pas construire une frontière qui se « rapproche » plus des données du premier groupe que des données du deuxième groupe. Comme il est difficile de voir en quoi deux mesures gaussiennes ayant des matrices de covariance de rang plein, peuvent (en dimension infinie) être orthogonales, nous donnons un exemple avec des mesures uniformes illustrant parfaitement nos propos.

Exemple 1.1 (Règle symétrique et données bien séparées)**.** *Supposons que dans le problème de classification à deux classes, $\theta \in]0, 1[$ détermine les lois $P_0 = \mathcal{U}[0, \theta]$ et $P_1 = \mathcal{U}[\theta, 1]$. Si θ est connu, on peut décider sans jamais se tromper qu'une observation X est issue de P_1 en regardant le signe de $X - \theta$. Si θ est inconnu, il existe de multiples manières d'apprendre le paramètre θ. Notons qu'aucune méthode ne permet, lorsque la taille de l'échantillon d'apprentissage est finie de produire une procédure qui ne fasse pas d'erreur de classification. Soient $X_1^i, \ldots, X_n^i$ les données d'apprentissage du groupe i $(i = 0, 1)$. La stratégie consistant à choisir $\hat{\theta}_- = \min_i X_i^1$ amène à sous-estimer θ et à favoriser les décisions prises dans le sens du groupe 0, à l'inverse choisir $\hat{\theta}_+ = \max_i X_i^0$ amène à favoriser les décisions prises dans le sens du groupe 1. Une meilleurs stratégie est d'utiliser une moyenne pondérée des deux estimateurs : $\hat{\theta} = \pi\hat{\theta}_+ + (1 - \pi)\hat{\theta}_-$. Par exemple, dans le cas où $\theta = 1/2$ et où $\pi = 1/2$, un tel estimateur permet d'avoir une erreur de première et de seconde espèce (sachant l'échantillon d'apprentissage) qui ont la même loi. On ne favorise alors aucun des deux groupes.*

[1] Par données bien séparées, nous voulons signifier l'existence d'un ensemble A tel que $P_0(A)$ est proche de 0 et $P_1(A)$ et proche de 1, autrement dit $d_1(P_1, P_0)$ proche de 1.

1.2.2 Deux résultats sur ce qui est à l'origine d'un mauvais apprentissage en grande dimension

Nous présentons deux résultats sur ce qui est à l'origine d'un mauvais apprentissage en grande dimension. La preuve de ces résultats est donnée au Chapitre 3 et repose sur le résultat fondamental suivant :

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \geq \frac{|\alpha|}{2\pi} e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}},$$

(cf dernier point du Théorème 3.1 Chapitre 3) où α est défini par (1.8).

Identification de la structure de correlation

Rappelons que si C est une matrice symétrique semi-définie positive, alors on peut définir son inverse généralisé, ou inverse de Moore Penrose : C^-. Cet inverse généralisé C^- est défini grâce à la décomposition $\mathbb{R}^p = Ker(C) \oplus Ker(C)^\perp$. Sur $Ker(C)$, C^- est nul, et sur $Ker(C)^\perp$, C^- et est égal à l'inverse de $\tilde{C} = C_{|Ker(C)^\perp}$ (i.e $\tilde{C}$ est la restriction de C à $Ker(C)^\perp$).

Proposition 1.1. *Supposons que l'échantillon d'apprentissage est constitué de $n < p$ observations et que μ_1 et μ_0 sont connus. Soient $\hat{C}$ la covariance empirique et $\hat{C}^-$ l'inverse généralisé de Moore-Penrose de $\hat{C}$. Alors en prenant $\hat{F}_{10} = \hat{C}^- m_{10}$ et $\hat{s}_{10} = s_{10}$, la règle de classification $1_{\hat{V}}$ définie par (1.7) vérifie*

$$\mathbb{E}_{P^{\otimes n}}[\mathcal{R}(\mathbb{1}_{\hat{V}})] \geq \frac{\arccos\left(\sqrt{\frac{n}{p}}\right)}{2\pi} e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}}.$$

La démonstration de cette proposition est donnée dans la Sous-section 3.1.4 du Chapitre 3 de cette partie. Commentons maintenant cette proposition.

Supposons que les estimateurs $\hat{F}_{10}$ et $\hat{s}_{10}$ utilisés sont construits à partir des estimateurs empiriques $(\bar{\mu}_k)_{k=0,1}$ et $\bar{C}$:

$$\hat{F}_{10} = \bar{C}^-(\hat{\mu}_1 - \hat{\mu}_0), \quad \hat{s}_{10} = (\bar{\mu}_1 + \bar{\mu}_0)/2. \tag{1.16}$$

La règle de décision correspondante s'appelle la règle de Fisher. La proposition précédente nous permet d'affirmer que, lorsque s_{10} est connu, la règle de Fisher donne une très mauvaise règle de classification en grande dimension. D'un point de vu heuristique, la raison à cela est que si la taille de l'échantillon d'apprentissage est $n \ll p$, l'estimateur empirique $\hat{C}$ de la matrice de covariance C est de rang n. Aussi, si l'on utilise l'inverse de Moore-Penrose associé à cet opérateur la seule information utilisée est alors concentrée dans un espace de dimension n choisi à peu près au hazard. Notons que Bickel et Levina [12] obtiennent un résultat asymptotique allant dans ce sens. Notre résultat est non-asymptotique et concerne l'erreur d'apprentissage.

Une alternative courante à l'inversion de $\bar{C}$ est de supposer qu'une observation X est composée d'attributs qui sont des réalisations de variables aléatoires réelles indépendantes. Ceci revient à supposer que les données sont observées dans une base qui diagonalise la matrice de covariance C. Sous cette hypothèse, on remplace $\bar{C}$ par $Diag(\bar{C})$, où $Diag$ est l'opérateur qui a une matrice A associe la matrice diagonale obtenue en fixant à zéro tous les éléments hors diagonal de A. Dans $\mathcal{X} = \mathbb{R}^p$, cette hypothèse peut être vérifiée par exemple lorsque l'on suppose que C est de type Toeplitz (i.e $C_{ij} = c(i-j)$ avec $c : \mathbb{Z} \to \mathbb{R}$ une suite p-périorique). Ces matrices correspondent

à des opérateurs de convolution circulaires et sont diagonales dans la base de Fourier discrète $(g^m)_{0\leq m<p}$ où

$$(g^m)_k = \frac{1}{\sqrt{p}} \exp\left(\frac{2i\pi mk}{p}\right).$$

Par exemple, Bickel et Levina [12], proposent d'utiliser la base de Fourier pour effectuer de la classification sur les processus stationnaires. A un processus stationnaire correspond une mesure gaussienne infinie dimensionnelle dont l'opérateur de covariance est diagonal dans la base de Fourier. Ce type de stratégie, lorsque la dimension p est grande ou tend vers l'infini, est bien connu dans les problèmes d'estimation de matrices de covariance (voir par exemple [54] ou [26] et les références qui y sont faites). Il s'agit, lorsque le nombre d'observations à disposition n'est pas suffisamment important, de faire des hypothèses statistiques permettant de réduire le nombre de paramètres à estimer, et d'allier estimation et approximation.

Notons aussi l'existence de démarches intermédiaires basées sur l'hypothèse d'une structure par bloc de la matrice de covariance (voir par exemple [11]). De manière générale, tout ce qui peut constituer à la fois une hypothèse interprétable par des propriétés statistiques, et une source de réduction du nombre de paramètres à estimer, est intéressant.

Sur l'estimation linéaire de F_{10}.

Proposition 1.2. *Supposons que C est une matrice définie positive, que l'échantillon d'apprentissage est constitué de n observations, que $\hat{F}_{10} = C^{-1}(\bar{\mu}_1 - \bar{\mu}_0)$ et que $\hat{s}_{10} = s_{10}$. Alors, la règle de classification $\mathbb{1}_{\hat{V}}$ définie par (1.7) vérifie*

$$\mathbb{E}_{P^{\otimes n}}[\mathcal{R}(\mathbb{1}_{\hat{V}})] \geq \frac{\arccos\left(\frac{1}{\sqrt{p-2}}(\sqrt{n}\|F_{10}\|_{L_2(\gamma_C)} + 1)\right)}{2\pi} e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}}.$$

Cette proposition est démontrée à la sous-section (3.1.5) du Chapitre 3.

Supposons qu'il existe $0 < r < R$ tels que $R > \|F_{10}\|^2_{L_2(\gamma_C)} \geq r$. D'après la proposition précédente, uniformément sur les valeurs possibles de μ_1 et μ_0, l'erreur d'apprentissage et l'excès de risque ne peuvent tendre vers 0 que si $\frac{n}{p}$ tend vers 0. Rappelons que si aucune hypothèse a priori n'est faite sur m_{10}, $\bar{m}_{10}$ est le meilleur estimateur (au sens des moindres carrés) de m_{10}. Ainsi, de la même manière que dans des problématiques d'estimation d'un vecteur en grande dimension telles que celles décrite dans ([19]), il faut faire une hypothèse plus restrictive sur m_{10}. Nous supposerons, dans le Chapitre 2 de la Partie II, que si $(a_k)_{k\geq 0}$ sont les coefficients de $C^{-1/2}m_{10}$ dans une base bien choisie, $\sum_{k\geq 0} a_k^q \leq R^q$ pour $0 < q < 2$.

1.2.3 Cas où $\|F_{10}\|_{L_2(\gamma_C)}$ diverge, données biens séparées.

Nous allons donner un résultat permettant de regarder le comportement limite de l'erreur d'apprentissage quand $\|F_{10}\|_{L_2(\gamma_C)}$ diverge. Nous allons donc étudier le comportement limite du problème lorsque la dimension p tend vers l'infini.

Theoreme 1.2. *Supposons que $0 < \alpha < \pi/2$ (α défini par l'équation (1.8)), et que $\cos(\alpha)\|F_{10}\|_{L_2(\gamma_C)} \to \infty$ quand p tend vers l'infini. Alors :*

$$\mathcal{R} \to \begin{cases} 0 & si\ \liminf_{p\to\infty} \frac{2|d_0|}{|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|} < 1 \\ b \geq \frac{1}{8} & si\ \limsup_{p\to\infty} \frac{2|d_0|}{|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|} > 1 \end{cases} \quad quand\ p \to \infty.$$

La preuve de ce théorème est donnée au Chapitre 3.

1.3 Perturbation quadratique d'une règle quadratique : QDA.

Dans toute cette section et dans la suite de ce mémoire, si $\mathcal{X}$ est un espace de Hilbert séparable et A un opérateur linéaire sur $\mathcal{X}$, A sera dit Hilbert Schmidt si pour une base orthonormée $(e_i)_i$ de $\mathcal{X}$, $\sum_i \|Ae_i\|^2_{\mathcal{X}} < \infty$. A tout opérateur Hilbert-Schmidt, on associe une norme Hilbert-Schmidt $\|A\|_{HS(\mathcal{X})} = \left(\sum_i \|Ae_i\|^2_{\mathcal{X}}\right)^{1/2}$. C'est un résultat classique d'algèbre que cette norme ne dépend pas de la base choisie.

Dans le cas où $C_1 \neq C_0$, $\mathcal{L}_{10}(x) = \mathcal{L}^Q_{10}(x)$ est un polynôme de degré deux sur $\mathbb{R}^p$:

$$\mathcal{L}^Q_{10}(x) = -\frac{1}{2}\langle A_{10}(x - s_{10}), x - s_{10}\rangle_{\mathbb{R}^p} + \langle G_{10}, x - s_{10}\rangle_{\mathbb{R}^p} - c, \tag{1.17}$$

où

$$A_{10} = C_1^{-1} - C_0^{-1}, \;\; G_{10} = S_{10}m_{10}, \tag{1.18}$$

$$S_{10} = \frac{C_0^{-1} + C_1^{-1}}{2}, \;\; c = \frac{1}{8}\langle Am_{10}, m_{10}\rangle_{\mathbb{R}^p} + \frac{1}{2}\log|\det(C_0^{-1}C_1)|,$$

m_{10} et s_{10} sont définis par (1.6). Nous donnons les calculs justifiant cette expression dans le Chapitre 1 de la Partie I.

Dans le problème de classification supervisée, l'échantillon d'apprentissage doit nous permettre de construire des estimateurs $\hat{G}_{10}$, $\hat{s}_{10}$ $\hat{A}_{10}$ et $\hat{c}$ de G_{10}, s_{10} A_{10} et c. Nous décidons que X est issu de P_1 s'il appartient à

$$\hat{V} = \left\{x \in \mathbb{R}^p \text{ tq } \widehat{\mathcal{L}}^Q_{10}(x) \geq 0\right\}, \tag{1.19}$$

où $\widehat{\mathcal{L}}^Q_{10}(x)$ est défini en substituant $\hat{G}_{10}$, $\hat{s}_{10}$ $\hat{A}_{10}$ et $\hat{c}$ à G_{10}, s_{10} A_{10} et c dans (1.17). Le théorème suivant donne le lien entre erreur d'estimation du logarithme du rapport de vraisemblance et l'erreur d'apprentissage.

Theoreme 1.3. *Soit γ une mesure gaussienne. Supposons que $\|\mathcal{L}^Q_{10}\|_{L_2(\gamma)} \geq r$. Alors, pour tout $q \in]0,1[$, il existe $c_1(r,q) > 0$ tel que*

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq c_1(r,q)\|\mathcal{L}^Q_{10} - \widehat{\mathcal{L}}^Q_{10}\|^{q/3}_{L_2(\gamma)}, \tag{1.20}$$

pour $\hat{V}$ donné par (1.19) et $\mathcal{R}$ défini par (1.5).

Ce théorème et sa version infinie-dimensionnelle sont une conséquence directe du point 1 du Théorème 3.2 Chapitre 3 Partie II. Ils ne résultent pas, comme dans le cas de la procédure LDA, de considérations géométriques, mais sont la conséquence d'un théorème général concernant les perturbations quadratiques de règles quadratiques. Ce théorème général est lui démontré par des techniques similaires à celles utilisées par Audibert et Tsybakov [6].

Commentaires.
Si l'on note

$$\delta_0 = \hat{c} - c + \left\langle \hat{G}_{10} + (\hat{A}_{10}^* + \hat{A}_{10})(\hat{s}_{10} - s_{10}), \hat{s}_{10} - s_{10} \right\rangle_{\mathbb{R}^p}, \tag{1.21}$$

(à une matrice A, on associe sa transposée A^*)

$$\delta^L = \hat{G}_{10} - G_{10} + (\hat{A}_{10}^* + \hat{A}_{10})(\hat{s}_{10} - s_{10}) \tag{1.22}$$

et

$$\delta^Q = \hat{A}_{10} - A_{10}, \tag{1.23}$$

on a alors par quelques calculs élémentaires :

$$\forall x \in \mathbb{R}^p \quad \widehat{\mathcal{L}}_{10}^Q(x) = \mathcal{L}_{10}^Q(x) + \delta_0 + \langle \delta^L, x - s_{10} \rangle_{\mathbb{R}^p} - \frac{1}{2}\langle \delta^Q(x - s_{10}), x - s_{10} \rangle_{\mathbb{R}^p}. \tag{1.24}$$

Aussi, nous parlerons de perturbation quadratique de règle quadratique. Nous déduisons du Théorème 1.3 le corollaire suivant.

Corollaire 1.1. *Soient $\mathcal{X} = \mathbb{R}^p$ et C une matrice symétrique définie positive sur $\mathbb{R}^p$. Supposons qu'il existe $r > 0$ tel que $\|\mathcal{L}_{10}\|^2_{L_2(\gamma_{C,s_{10}})} > r$. Alors, pour $\mathbb{1}_{\hat{V}}$ donné par (1.19) et quel que soit $0 < q < 1$ il existe une constante $c_1(r,q)$ telle que :*

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq c_1(r,q) \left(\frac{1}{2}\|C(A_{10} - \hat{A}_{10})\|^2_{HS(\mathbb{R}^p)} + \|C^{1/2}\delta^L\|^2_{\mathbb{R}^p} + 2\delta_0^2 + \frac{1}{2} trace^2(C(A_{10} - \hat{A}_{10})) \right)^{q/3},$$

où $\|A\|_{HS}$ est la norme Hilbert-Schmidt de la matrice A, $trace(A)$ est la trace de A (somme des éléments diagonaux de A) , δ^L est donné par (1.22) et δ_0 par (1.21).

Démonstration. Rappelons que δ^Q est donné par (1.23) et que

$$q_{\delta^Q}(x) = \langle \delta^Q x, x \rangle_{\mathbb{R}^p}.$$

On a

$$\begin{aligned}
\|\mathcal{L}_{10} - \widehat{\mathcal{L}}_{10}\|^2_{L_2(\gamma_{C,s_{10}})} &= \|\frac{1}{2}(\delta^Q(x) - \mathbb{E}_{\gamma_C}[\delta^Q(X)]) - \langle \delta^L, x \rangle_{\mathbb{R}^p} - (\delta_0 - \frac{1}{2}\mathbb{E}_{\gamma_C}[q_{\delta^Q}(X)])\|^2_{L_2(\gamma_C)} \\
&\leq \frac{1}{4} Var(q_{C^{1/2}\delta^Q C^{1/2}}(\xi)) + Var(\langle C^{1/2}\delta^L, \xi \rangle_{\mathbb{R}^p}) + 2\delta_0^2 + 2\mathbb{E}^2_{\gamma_C}[q_{C^{1/2}\delta^Q C^{1/2}}(\xi)] \\
&\quad (\xi \rightsquigarrow \gamma_{I_p,0}, \text{ notons qu'il y a en fait ici égalité}) \\
&= \frac{1}{2}\|C^{1/2}\delta^Q C^{1/2}\|^2_{HS(\mathbb{R}^p)} + \|C^{1/2}\delta^L\|^2_{\mathbb{R}^p} + 2\delta_0^2 + \frac{1}{2} trace^2(C^{1/2}\delta^Q C^{1/2}).
\end{aligned}$$

□

L'expression (1.18) de $\mathcal{L}_{10}^Q(x)$ peut être modifiée en utilisant le fait que

$$A_{10} = \frac{1}{2}\left(C_1^{-1/2} W_{10} C_1^{-1/2} - C_0^{-1/2} W_{01} C_0^{-1/2} \right) \text{ où } W_{ij} = I - C_i^{1/2} C_j^{-1} C_i^{1/2}. \tag{1.25}$$

Cette modification a un double intérêt. Elle fait intervenir W_{ij} qui, nous le verrons par la suite, tend à être (lorsque la dimension p grandit) un opérateur symétrique dont les valeurs propres forment une suite de l^2. Une telle suite est plus facile à estimer qu'une suite de l^∞. Elle fait intervenir W_{10} autant que W_{01} et si les méthodes pour estimer ces matrices sont symétriques par rapport aux échantillons d'apprentissage des deux groupes, la règle résultante sera symétrique au sens de la Définition 1.1.

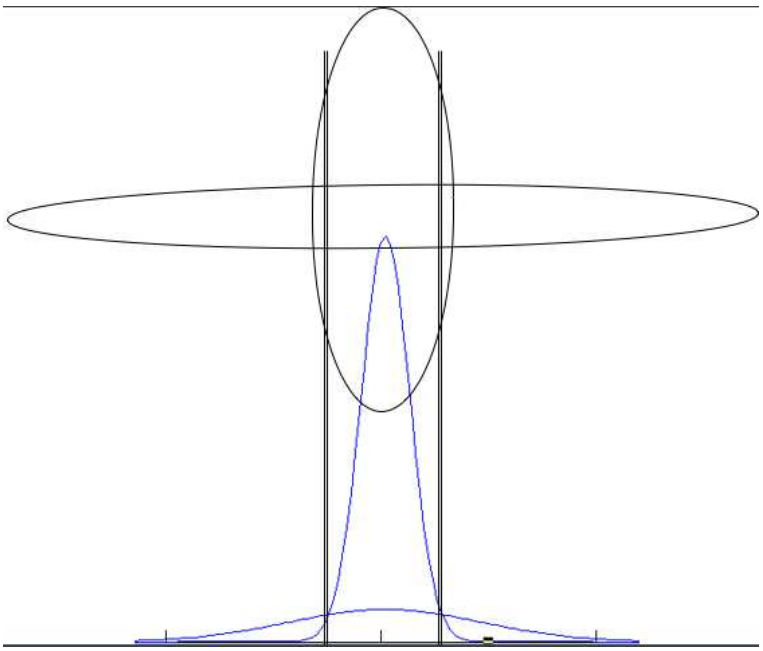


FIG. 1.2 – Séparation des données dans une direction ou les variances sont nettement différentes. Les deux groupes sont matérialisés par deux ellipsoïdes de concentration : l'une horizontale et l'autre verticale. Les deux groupes ont la même moyenne, c'est leurs structures de covariance qui diffèrent et c'est cela qui provoque la séparation des données. On ne peut tirer partie de cette séparation que si l'on utilise une règle quadratique.

Comparaison de ces résultats avec ceux de la procédure LDA. Ces résultats sont moins forts et moins précis que ceux obtenus pour la procédure LDA et certaines conjectures peuvent être faites dans un parallèle avec le Théorème 1.1. Dans ce théorème concernant les règles linéaires et dans le Théorème 1.2, nous avons expliqué en quoi les erreurs d'estimation des paramètres sont moins importantes lorsque $\|F_{10}\|_{L_2(\gamma_C)}$ est grand. Cette observation était basée sur la présence d'un terme qui décroît exponentiellement avec $\|F_{10}\|_{L_2(\gamma_C)}$ dans les quantités bornant l'erreur d'apprentissage. Dans le Théorème 1.3 concernant la procédure QDA, ce type de terme n'a pas été obtenu. Pourtant nous pensons qu'un tel terme existe. Pour donner un poids à cette conjecture, nous verrons qu'il est plus pertinent de la présenter dans le cadre infini-dimensionnel que nous allons donner à la Section 4.

Il s'agit par ailleurs de clarifier l'hypothèse consistant à supposer que $\mathcal{L}_{10}^{Q}$ est bornée inférieurement. Rappelons que cette hypothèse garantie que la contante c_1 dans l'équation (1.20) est indépendante des paramètres du problème. Dans le parallèle avec la procédure LDA, supposer que $\|\mathcal{L}_{10}^{Q}\|_{L_2(\gamma)}$ est borné inférieurement reviens à faire l'hypothèse que les lois des deux groupes sont tout le temps discernables. Cette hypothèse reflète essentiellement une caractéristique de la mesure d'erreur choisie : lorsque les mesures des deux groupes sont à peu près égales, le problème est rendu difficile par l'erreur d'apprentissage alors que toutes les règles se valent (cf Chapitre 5 Partie I). On peut bien évidement la présenter comme une hypothèse écartant un cas trivial dans la pratique.

Nous ne revenons pas ici sur l'estimation de G_{10} qui relève des mêmes problèmes que ceux rencontrés pour l'estimation de F_{10} dans le cadre de la procédure LDA. Nous allons maintenant examiner l'estimation de W_{01} (et donc de A_{10}) et de sa quantité symétrique W_{10}.

Problématique commune à l'estimation d'opérateur et à la classification : linéarisation de procédure. Rappelons que W_{10} est une matrice symétrique. Supposons que l'on connaît une base orthonormale dans laquelle elle est diagonale. Notons $\lambda_{10} = (\lambda_{10i})_{i=1,\dots,p}$ le vecteur de ses valeurs propres. Pour construire $\hat{W}_{10}$, il faut estimer ses valeurs propres et mesurer l'erreur d'estimation en norme l^2. Supposons que p tende vers l'infini. Nous verrons par la suite que si les mesures des groupes 0 et 1 tendent à être équivalentes, alors W_{10} tend à être un opérateur Hilbert-Schmidt. L'erreur d'estimation en norme l^2 des valeurs propres est une erreur d'estimation de W_{10} en norme Hilbert-Schmidt : $\|W_{10} - \hat{W}_{10}\|_{HS(\mathcal{X})}$. Encore une fois, si le vecteur λ_{10} est creux (ses coefficients décroissent assez vite), l'estimation par seuillage est souhaitable. Dans le cadre de la classification, ce seuillage ne correspond pas à une réduction de dimension mais à une linéarisation. En effet, si $\hat{W}_{10} = \sum_{i=1}^{l} \hat{\lambda}_{10i} e_i \otimes e_i$ pour $l \leq p$ et $(e_i)_{i=1,\dots,p}$ une base orthonormale de $\mathbb{R}^p$, on a :

$$\widehat{\mathcal{L}}_{10}^{Q} = \sum_{i=1}^{l} \hat{\lambda}_{10i} \langle e_i, x - \hat{s}_{10} \rangle_{\mathbb{R}^p}^2 + g(x),$$

où $g(x)$ est une application affine définie sur $\mathbb{R}^p$. Dans ce cas la règle de décision est affine dans un sous-espace de dimension $p - l$ et quadratique dans le sous-espace de dimension l engendré par $(e_i)_{i=1,\dots,l}$.

Notons que puisque $W_{10} = I - C_1^{-1/2} C_0 C_1^{-1/2}$, fixer les valeurs propres de $\hat{W}_{ij}$ à zéro dans un sous-espace de $\mathbb{R}^p$, revient à décider que dans ce sous espace les matrices de covariances C_1 et C_0 sont « assez proches » . Dans ce sous-espace on peut donc supposer que les matrices de covariances C_1 et C_0 sont égales. La règle de classification y est linéaire. La Figure 1.2 illustre le cas où les valeurs propres de W_{10} sont grandes et en quoi une règle quadratique est souhaitable dans ce cas.

L'estimation par seuillage des paramètres de la procédure LDA nous amenait à parler de réduction de dimension. Dans le cadre de la procédure QDA, l'estimation par seuillage de G_{10} constitue une réduction de dimension. L'estimation de W_{10} et W_{01} par seuillage induit la transformation d'une règle quadratique en une règle linéaire dans un sous espace donné. Ceci constitue une simplification de la règle.

1.4 Cas de données fonctionnelles

1.4.1 Motivation

Le problème médical comporte des données dont la dimension est en général de 256 ou 1024, mais les échantillons d'apprentissage comprennent entre 10 et 30 courbes par type de tissu cancéreux. Nous avons indiqué jusqu'ici comment construire une méthode qui soit justifiable lorsque la dimension p de l'espace dans lequel les courbes sont observées est grand par rapport au nombre n d'observations. Pourtant il reste encore quelques zones d'ombres à notre tableau parmi lesquelles les suivantes seront éclaircies.

1. Nous avons effectué certains passages à la limite (notamment dans le Théorème 1.2). Il est intéressant de comprendre quelle est la nature des objets mathématiques utilisés lorsque la dimension p tend vers l'infini.
2. Dans le cas de la règle linéaire, les Théorèmes 1.1 et 1.2 permettent de mettre en évidence le comportement d'une règle de décision lorsque la dimension grandie et que les mesures $\gamma_{\mu_0,C}$ et $\gamma_{\mu_1,C}$ tendent à être orthogonales (quand $\|F_{10}\|_{L(\gamma)}$ tend vers l'infini). Dans ce cas, nous avons clairement établit que le problème d'apprentissage tend à être plus simple. Dans le Théorème 1.3 concernant la règle quadratique nous ne sommes pas parvenu à un tel constat.
3. Dans le cas de la règle quadratique, d'autres problèmes persistent. Par exemple, l'opérateur $A_{10} = C_1^{-1} - C_0^{-1}$ qui définit la partie quadratique de $\mathcal{L}_{10}$ est a priori un opérateur qui, lorsque la dimension de l'espace va grandir, ne sera pas nécessairement borné. En effet un opérateur de covariance associé à une mesure gaussienne dans un espace de Hilbert de dimension infinie est forcement compact et a donc des valeurs propres qui s'accumulent en 0. Autrement dit les valeurs propres de A_{10} formeront une suite différence de deux suites dont les termes tendent vers l'infini. Il est donc a priori difficile d'espérer une bonne estimation de cette matrice en grande dimension.

Les objets infinis dimensionnels dont nous rappelons les définitions et les propriétés dans l'Annexe B, vont nous aider à mieux quantifier certains enjeux de la classification en grande dimension. Si $\mathcal{X}$ est un espace de Banach, γ une mesure gaussienne sur $\mathcal{X}$, nous noterons $\mathcal{X}^*$ le dual topologique de $\mathcal{X}$. Nous allons utiliser les espaces suivant qui sont définis dans l'Annexe B. L'espace auto reproduisant $H(\gamma)$, l'espace des formes affines mesurables d'intégrale nulle par rapport à γ, noté $\mathcal{X}^*_\gamma$, l'espace des polynômes mesurables de degrés au plus deux et de carré intégrable $\mathcal{X}^*_{2,\gamma}$, l'espace des polynômes mesurables de degrés deux, d'intégrale nulle et de carré intégrable par rapport à γ : $E_2(\gamma)$. Nous avons dans $L_2(\gamma)$:

$$\mathcal{X}^*_{2,\gamma} = \{ \text{ applications constantes } \} \oplus \mathcal{X}^*_\gamma \oplus E_2(\gamma).$$

Notons seulement ici que d'une part les éléments de $\mathcal{X}^*_\gamma$ nous permettent de décrire lorsque p tend vers l'infini, le type d'objet vers lequel $\mathcal{L}^L_{10}$ (défini par (1.6)) tend γ-presque sûrement, et que d'autre part, les éléments de $\mathcal{X}^*_{2,\gamma}$ nous permettent de décrire le type d'objet vers lequel $\mathcal{L}^Q_{10}$ (défini par (1.17)) tend γ-presque sûrement.

Pour les mesures gaussiennes, considérer des données infinies dimensionnelles permet de tirer partie d'un résultat essentiel : deux mesures gaussiennes sont soit équivalentes soit orthogonales (cf Chapitre 1 Partie I). Nous allons voir en quoi cette dichotomie, et surtout ce qui la caractérise, est à l'origine d'un certain nombre d'explications et de conjectures formulées.

1.4.2 Problème de détection en dimension infinie et perturbation : travaux existants

Notre travail, dans le cadre infini-dimensionnel, a eu pour point de départ le problème de détection initié par Grenander [37], Rao et Varadarajan [62] . Nous rappelons qu'il s'agit, au

vu de l'observation de X, une courbe aléatoire (un élément de $\mathcal{X}$), de tester les hypothèses (les mêmes que celles définies par (1.1)) :

$$H_0 \ : \ X \rightsquigarrow P_0 = \gamma_{C_0,\mu_0} \text{ contre } H_1 \ : \ X \rightsquigarrow P_1 = \gamma_{C_1,\mu_1},$$

où γ_{C_0,μ_0} et γ_{C_1,μ_1} sont deux mesures gaussiennes sur $\mathcal{X}$ un espace de Hilbert séparable. Les opérateurs de covariance sont supposés d'images denses dans $\mathcal{X}$. Si $\mathcal{X}$ est un espace de Hilbert séparable, les deux mesures gaussiennes considérées sont équivalentes si et seulement si (cf [15])

$$m_{10} = \mu_1 - \mu_0 \in H(\gamma_{C_1,\mu_1}) = H(\gamma_{C_0,\mu_0}), \tag{1.26}$$

et

$$W_{10} = I - C_1^{1/2} C_0^{-1} C_1^{1/2} \in HS(\mathcal{X}), \tag{1.27}$$

où $HS(\mathcal{X})$ est l'ensemble des opérateurs Hilbert Schmidt sur $\mathcal{X}$, et $H(\gamma_{C_1,\mu_1})$ est l'espace de Hilbert Reproduisant de la mesure gaussienne γ_{C_1,μ_1}.

Dans le cas où ces lois sont équivalentes, la dérivée de radon nykodim $f(x) = \frac{dP_1}{dP_0}(x)$ permet de définir un test de Neymann Pearson. Lorsque l'on cherche à minimiser la somme des erreurs de première et de seconde espèce, ce test correspond à rejeter H_0 si X appartient à

$$V = \{x \in \mathcal{X} \ : \ \mathcal{L}_{10}(x) \geq 0\} \quad \text{où } \mathcal{L}_{10}(x) = \log(f(x)).$$

On peut encore donner un sens à $\mathcal{L}_{10}^A$ et $\mathcal{L}_{10}^Q$ définis respectivement par (1.6) et (1.17) en remplaçant le produit scalaire de $\mathbb{R}^p$ par celui de $\mathcal{X}$ (ou par le produit de dualité de $\mathcal{X}$ dans le cas d'un espace de Banach) et en passant à la limite sur la dimension.

Nous allons montrer que les Théorèmes 1.1 et 1.3 ainsi que le Corollaire 1.1 restent valables si $\mathcal{X}$ n'est plus $\mathbb{R}^p$ mais un espace de Hilbert séparable (en fait nous montrons que les Théorèmes 1.1 et 1.3 sont encore valables dans un espace de Banach séparable).

Notons qu'il existe une littérature associée à l'étude de la stabilité dans les problèmes de détection de signal (voir par exemple [38] ou [44] et les références qui y sont faites). Dans ces approches, une (ou deux) hypothèse(s) P est contaminée. Autrement dit, elle est remplacée par une loi de la forme $(1-\epsilon)P + \epsilon Q$. Ici, ce sont les paramètres qui sont perturbés et cette démarche n'a pas encore été envisagée dans le cas de la dimension infinie ou finie.

1.4.3 Procédure LDA dans un espace de Banach séparable.

Nous allons donner les règles infinies dimensionnelles et les résultats obtenus dans le cas d'espaces de Banach pour la procédure LDA. Le cas de la procédure QDA sera traité à la sous-section suivante.

Supposons que $\mathcal{X}$ est un espace de Banach séparable, que pour $k = 0, 1$ $P_k = \gamma_{C,\mu_k}$ et que P_1 et P_0 sont équivalentes, c'est-à-dire que $m_{10} = \mu_1 - \mu_0 \in H(\gamma_{C,\mu_1}) = H(\gamma_{C,\mu_0})$. La régle de Neymann Pearson associe à $x \in \mathcal{X}$ le groupe 1, si $f(x) \geq 0$, où

$$f(x) = F_{10}(x) \text{ avec } s_{10} = \frac{\mu_0 + \mu_1}{2} \tag{1.28}$$

et F_{10} est l'unique élément de $\mathcal{X}^*_{\gamma_{C,s_{10}}}$ associé à $m_{10} \in H(\gamma_{C,s_{10}})$ (voir Annexe B). Notons que F_{10} est défini ici de manière abstraite. Du point de vu applicatif, il est plus intéressant de voir cette notation en terme de limite : il existe une suite d'éléments $(C^{-1}m_{10}^p)_{p\in\mathbb{N}}$ de $\mathcal{X}^*$ tels que $\gamma_{C,s_{12}}$-presque sûrement $F_{10}(x) = \lim_{p\to\infty}\langle C^{-1}m_{10}^p, x - s_{10}\rangle_{\mathcal{X}^*,\mathcal{X}}$.

Remarque 1.1. *Nous avons écrit, dans le cadre de la dimension finie, que $\|F_{10}\|_{L_2(\gamma_C)}$ mesurait la séparation des données. Si les mesures γ_{C,μ_1} et γ_{C,μ_0} sont équivalentes, on a*

$$\|\mathcal{L}_{10}^L\|_{L_2(\gamma_{C,s_{10}})} = \|F_{10}\|_{L_2(\gamma_{C,s_{10}})} = \|m_{10}\|_{H(\gamma_{C,s_{10}})}.$$

Par ailleurs, si $\|F_{10}\|_{L_2(\gamma_{C,s_{10}})} \to \infty$, alors les mesures correspondantes γ_{C,μ_1} et γ_{C,μ_0} tendent à être orthogonale. On parle alors de séparation parfaite des données (voir Chapitre 1 de la Partie I).

De même qu'en dimension finie, si $\hat{V}$ est une partie mesurable de $\mathcal{X}$, on définit l'erreur d'apprentissage par

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) = \frac{1}{2}\left(P_0\left(X \in \hat{V} \setminus V\right) + P_1\left(X \in V \setminus \hat{V}\right)\right). \tag{1.29}$$

On a alors le théorème suivant, qui est la version infinie-dimensionnelle du Théorème 1.1.

Theoreme 1.4. *Soient $\hat{s}_{10} \in \mathcal{X}$, $\hat{F}_{10} \in \mathcal{X}^*_{\gamma_{C,s_{10}}}$, $\mathcal{L}_{10}(x) = \hat{F}_{10}(x) - \hat{F}_{10}(\hat{s}_{10})$, et*

$$\hat{V} = \{x \in \mathcal{X} \ : \hat{\mathcal{L}}_{10}(x) \geq 0\}. \tag{1.30}$$

Alors, on a l'inégalité suivante :

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_{C,s_{10}})}} \quad \text{où } \mathcal{E} = \left(\frac{4\|F_{10}\|_{L_2(\gamma_{C,s_{10}})}}{\sqrt{\pi}\|\hat{F}_{10}\|_{L_2(\gamma_{C,s_{10}})}}|d_0| + \|F_{10} - \hat{F}_{10}\|_{L_2(\mathcal{X},\gamma_{C,s_{10}})}\right),$$

où $d_0 = -\hat{F}_{10}(\hat{s}_{10})$.
De plus, si $|d_0| \leq \frac{1}{4}|\langle\hat{F}_{10}, F_{10}\rangle_{L_2(\gamma_{C,s_{10}})}|$ et $\langle\hat{F}_{10}, F_{10}\rangle_{L_2(\gamma_{C,s_{10}})} \geq \frac{\sqrt{2}}{2}\|F_{10}\|_{L_2(\gamma_{C,s_{10}})}\|\hat{F}_{10}\|_{L_2(\gamma_{C,s_{10}})}$, alors

$$\mathcal{R}(1_{\hat{V}}) \leq e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_{C,s_{10}})}}{32}}\frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_{C,s_{10}})}}. \tag{1.31}$$

Démonstration. La démonstration consiste en un simple passage à la limite sur la dimension couplée avec l'utilisation de la configuration cylindrique de la mesure gaussienne. Afin de ne pas alourdir les notations, nous ne mentionnerons pas ici l'indice « 10 » . Comme nous l'avons déjà dit, une propriété essentielle de la mesure gaussienne $\gamma_{C,s}$ est que son comportement peut se décrire comme limite de ses restrictions à des sous-espaces de dimensions finies. Nous allons utiliser cette propriété. Soient $(e_i)_{i\in\mathbb{N}}$ une base orthonormée de $H(\gamma_{C,s})$, $(e_i')_{i\in\mathbb{N}}$ la base de $\mathcal{X}^*_{\gamma_{C,s}}$ associée (voir Annexe B) et $P_p \ : \ x \in \mathcal{X} \to \sum_{i=1}^p e_i'(x)e_i$, $m \in H(\gamma_{C,s})$. Nous rappelons (voir aussi Annexe B) que $P_p(x)$ permet de ramener l'intégrale d'une fonction $f \in L^1(\gamma_{C,s})$, à la limite de l'intégrale de f par rapport $\gamma_{C_p,s_p} = \gamma \circ P_p^{-1}$ lorsque p tend vers l'infini. En effet, on a :

$$\lim_{p\to\infty}\int_{\mathcal{X}} f(x)\gamma_{C_p,s_p}(dx) = \lim_{p\to\infty}\int_{\mathcal{X}} f(P_p(x))\gamma_{C,s}(dx)$$

ce qui d'après le théorème de convergence dominée de Lebesgue permet d'écrire :

$$\lim_{p\to\infty}\int_{\mathcal{X}} f(x)\gamma_{C_p,s_p}(dx) = \int_{\mathcal{X}} f(x)\gamma_{C,s}(dx). \tag{1.32}$$

Pour utiliser cette représentation cylindrique, nous noterons dans la suite

$$m_p = \sum_{i=1}^{p}\langle m, e_i\rangle_{H(\gamma_{C,s})}e_i \ \ , \ \ F_p(x-s_p) = \sum_{i=1}^{p}\langle F, e_i'\rangle_{L_2(\gamma_{C,s})}e_i'(x),$$

$$d_p(x-s_p) = \sum_{i=1}^{p}\langle d, e_i'\rangle_{L_2(\gamma_{C,s})}e_i'(x).$$

Attention $F_p(x)$ et $d_p(x)$ sont d'intégrale nulle par rapport à la mesure γ_C et convergent γ_C-presque sûrement vers $F(x+s)$ et $d(x+s)$. Retenons surtout que ces applications sont linéaires et qu'elles peuvent être assimilées à des formes linéaires sur $\mathbb{R}^p$ et donc à des vecteurs de $\mathbb{R}^p$. Ces notations permettent de tirer parti de l'équation (1.32). D'une part, si l'on note

$$\mathcal{R}_p = \frac{1}{2}\gamma_{C_p,s_p}\left((V_{\langle x-s_p,F_p\rangle_{\mathbb{R}^p}} \setminus V_{\langle x-s_p,F_p+d_p\rangle_{\mathbb{R}^p}+d_0}) + m_p\right)$$

$$+\frac{1}{2}\gamma_{C_p,s_p}\left((V_{\langle x-s_p,F_p+d_p\rangle_{\mathbb{R}^p}+d_0} \setminus V_{\langle x-s_p,F_p\rangle_{\mathbb{R}^p}}) - m_p\right),$$

on a

$$\lim_{p\to\infty}\mathcal{R}_p(\mathbb{1}_{\hat{V}}) = \mathcal{R}(\mathbb{1}_{\hat{V}}), \tag{1.33}$$

et d'autre part, lorsque p tend vers l'infini

$$\forall f \in \mathcal{X}^*_{\gamma_{C,s}} \quad \|f\|_{L_2(\gamma_{C_p,s_p})} \to \|f\|_{L_2(\gamma_{C,s})}$$

ce qui implique les limites sur la borne supérieure finie dimensionnelle du Théorème 1.1 □

1.4.4 Deuxième théorème : analyse de la QDA dans un espace de Hilbert

Soient $\mathcal{X}$ un espace de Hilbert séparable, γ_{C_1,m_1} et γ_{C_0,m_0} deux mesures gaussiennes équivalentes sur $\mathcal{X}$ de covariances respectives C_1 et C_0 avec $Im(C_0)$ et $Im(C_1)$ denses dans $\mathcal{X}$ et de moyennes m_1 et m_0. Nous noterons $(\lambda_i)_{i\geq 1}$ les valeurs propres de W_{10} (défini par 1.27). On a (cf [15], p293) :

$$\mathcal{L}^Q_{10} = \frac{1}{2}q^{\gamma_{C_1,s_{10}}}_{W_{10}}(x) + G_{10}(x) + \sum_{i=1}^{\infty}\left(\lambda_i - \frac{1}{2}\log(\frac{1}{1-\lambda_i})\right), \tag{1.34}$$

où $G_{10} = \frac{G_1+G_0}{2}$, G_1 et G_0 sont les éléments respectivement de $\mathcal{X}^*_{\gamma_{C_1,s_{10}}}$ et $\mathcal{X}^*_{\gamma_{C_0,s_{10}}}$ associés tous deux à $m_{10} \in H(\gamma_{C_1,s_{10}}) = H(\gamma_{C_0,s_{10}})$ et $q^{\gamma_{C_1,s_{10}}}_{W_{10}}$ est l'élément de $E_2(\gamma_{C_1,s_{10}})$ associé à W_{10} (voir Annexe B pour sa construction). Notons que cette écriture est une version limite de celle donnée par l'équation (1.17) en dimension finie. Nous avons le théorème suivant, qui est la version infinie-dimensionnelle du Théorème 1.3.

Theoreme 1.5. *Soit γ une mesure gaussienne sur $\mathcal{X}$. Supposons que $\|\mathcal{L}_{10}^{Q}\|_{L_2(\gamma)} \geq r$. Pour tout $0 < q < 1$, ll existe une constante $c_1(r,q)$ positive telle que pour tout $\widehat{\mathcal{L}}_{10}^{Q} \in \mathcal{X}_{2,\gamma}^{*}$ (voir annexe B définition B.1 pour la définition de $\mathcal{X}_{2,\gamma}^{*}$) :*

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq c_1(r,q)\|\mathcal{L}_{10}^{Q} - \widehat{\mathcal{L}}_{10}^{Q}\|_{L_2(\mathcal{X},\gamma)}^{q/3}, \tag{1.35}$$

où $\hat{V} = \{x \in \mathcal{X} \ : \ \widehat{\mathcal{L}}_{10}^{Q} \geq 0\}$.

Démonstration. Ce résultat est une conséquence directe du point 1 du théorème 3.2 Chapitre 3 Partie II. □

1.5 Perspectives

Les extensions des résultats obtenus pourraient être de plusieurs natures. Si l'on note ν une mesure de probabilité sur $\mathcal{X} = \mathbb{R}^p$ (ou un espace plus abstrait),

$$V = \{x \in \mathcal{X} \ : \ f(x) \geq 0\} \ \text{ et } \hat{V} = \{x \in \mathcal{X} \ : \ f(x) + \delta(x) \geq 0\},$$

Afin de borner supérieurement l'erreur d'apprentissage. on peut vouloir chercher à calculer des bornes inférieures ou supérieures (ce deuxième cas étant plus intéressant pour notre problème) de $\nu(V \setminus \hat{V})$ ou des bornes supérieures pour $\nu(V \triangle \hat{V})$[2] .

Les méthodes utilisées pour la démonstration des résultats concernant la procédure LDA sont géométriques et sont donc adaptés à des cas ou les ensembles V et $\hat{V}$ n'ont pas une interprétation géométrique trop complexe. Mais d'autres cas pourraient être envisagés. En effet, pour la démonstration de la borne supérieure du Théorème 1.1, nous utilisons essentiellement le fait que la mesure gaussienne centrée réduite est invariante par rotation et a une densité de la forme $e^{-\phi(x)}$ où ϕ est strictement convexe dans le sens ou il existe $c > 0$ tel que pour tout $x, y \in \mathbb{R}^n$

$$\phi(x) + \phi(y) - 2\phi\left(\frac{x+y}{2}\right) \geq \frac{c}{2}\|x - y\|_{\mathbb{R}^n}^2.$$

L'invariance par rotation pourrait être remplacée par une hypothèse n'autorisant pas la mesure considérée à charger une portion angulaire particulière.

Les méthodes utilisées pour la démonstration des résultats concernant la procédure QDA sont beaucoup plus générales. En particulier, elles résultent d'inégalités de grande déviation sur la perturbation δ. Ces inégalités existent pour d'autres types de mesures que la mesure gaussienne (voir par exemple [50]) et pour d'autres types d'applications que les applications quadratiques et linéaires. Dans le cas gaussien, par exemple, les bornes utilisées existent pour des polynômes mesurables de degrés finis. La démonstration donnée dans la Section 3 du Chapitre 3 est partagée de manière à ce que les outils qui pourraient servir à d'éventuelles généralisations soient facilement utilisables. Enfin, nous pensons qu'une utilisation probabiliste de la formule de Coaire (voir [29])

[2] Rappelons que $V \triangle \hat{V}$ est la différence symétrique entre les parties $\hat{V}$ et V, c'est-à-dire l'ensemble des éléments qui sont dans V mais pas dans $\hat{V}$ ou dans $\hat{V}$ mais pas dans V.

Chapitre 2

Méthodes de réduction de dimension pour la classification

> Ne demande pas que les événements arrivent comme tu veux, contente toi de les vouloir comme ils arrivent
>
> Epictete

Dans ce chapitre, nous donnons deux procédures de réduction de dimension. La première est destinée à un algorithme de classification et la deuxième à un algorithme de segmentation. Nous appliquons la procédure de classification à des données réelles.

2.1 Introduction

Dans ce chapitre nous allons donner une méthode de classification de données gaussiennes en grande dimension. Nous traitons le cas de plusieurs classes dans la pratique et donnons des résultats théoriques dans le cas de deux classes. Nous supposons que pour $k \in \{1, \dots, K\}$ et observons n_k vecteurs de dimension p : $(X_{ik})_{i=1,\dots,n_k}$. Ces observations constituent l'échantillon d'apprentissage. Nous noterons $n = \sum_{k=1}^{K} n_k$. Nous supposons que chacun des n_k vecteurs du groupe k est composé des p premiers coefficients d'ondelette (voir Annexe A) d'une courbe aléatoire de $\mathcal{X} = L^2[0,1]$ issue d'une loi gaussienne de moyenne et de covariance inconnues. En d'autres termes les données de l'échantillon d'apprentissage sont observées dans un sous-espace E_p de $\mathcal{X}$ de dimension finie p, engendré par les p premiers vecteurs de la base d'ondelette et assimilé à $\mathbb{R}^p$. Ainsi nous observons, dans l'échantillon d'apprentissage, des vecteurs aléatoires de loi $P_k = \gamma_{C_k,\mu_k}$ (sur $\mathbb{R}^p$) de moyenne et de covariance inconnues. Dans le contexte médical, chacune des lois $(P_k)_{k=1,\dots,K}$ modélise la distribution d'un spectre associé à un type histopathologique donné.

Nous allons donner deux méthodes de réduction de dimension utilisant l'échantillon d'apprentissage. La première aura pour objectif la construction d'une partition de $\mathbb{R}^p$ en $\hat{V}_1, \dots, \hat{V}_K$. Cette partition permettra de définir la règle de classification utilisée. Si une observation $X \in \mathbb{R}^p$ est issue d'une des K lois $P_1, \dots, P_K$ nous déciderons que X appartient à la classe k si $X \in \hat{V}_k$. La deuxième méthode de réduction de dimension servira à la construction d'estimateur des densités des lois P_k à $\mathbb{R}^p$. Ces estimateurs seront utilisés dans la troisième partie de ce mémoire pour la

segmentation d'images hyperspectrales.

Dans la première méthode nous construirons la partition $\hat{V}_1, \ldots, \hat{V}_K$ de $\mathbb{R}^p$ à partir de $\frac{K(K-1)}{2}$ fonctions frontières $(\widehat{\mathcal{L}}_{k_1k_2})_{1\leq k_1<k_2\leq K}$:

$$\hat{V}_k = \left\{ x \in \mathbb{R}^p \ : \ \forall j \in \{1, \ldots, K\} \ \ \widehat{\mathcal{L}}_{kj}(x) \geq 0 \right\}, \tag{2.1}$$

où $\widehat{\mathcal{L}}_{jj} = 0$, $\widehat{\mathcal{L}}_{k_1k_2} = -\widehat{\mathcal{L}}_{k_2k_1}$ si $k_2 < k_1$ et $\widehat{\mathcal{L}}_{ij}$ est un estimateur du logarithme du rapport de vraisemblance entre P_i et P_j.

Dans la deuxième méthode nous construirons des estimateurs $\hat{f}_1, \ldots, \hat{f}_K$ des K densités $f_1, \ldots, f_K$ des lois $P_1, \ldots, P_K$ à $\mathbb{R}^p$. Ces estimateurs doivent permettre de construire la partition $\hat{V}_1, \ldots, \hat{V}_K$ de $\mathbb{R}^p$:

$$\hat{V}_k = \left\{ x \in \mathbb{R}^p \ : \ \forall j \in \{1, \ldots, K\} \ \ \hat{f}_j(x) \leq \hat{f}_k(x) \right\}, \tag{2.2}$$

associée à une règle de classification performante.

2.2 Premier problème : estimation des fonctions frontières.

Nous divisons notre présentation en deux sous-sections. Dans la première, nous donnons des résultats théoriques dans le cas où les matrices de covariances sont supposées connues. Dans la deuxième, nous donnons la méthode utilisée lorsque la covariance est inconnue. Nous gardons les notations du chapitre précédent. Dans toute la suite de ce chapitre, $i, j \in \{1, \ldots, K\}$. Dans le cas de la procédure LDA, $m_{ij} = \mu_i - \mu_j$ $F_{ij} = C^{-1}m_{ij}$, $s_{ij} = \frac{\mu_i+\mu_j}{2}$, et dans le cas de la procédure QDA, $G_{ij} = \frac{1}{2}(C_i^{-1} + C_j^{-1})m_{ij}$, $A_{ij} = C_i^{-1} - C_j^{-1}$.

2.2.1 Cas de covariances connues et égales

Hypothèses et notations. Dans cette sous-section, $\bar{\mu}_k$ est la moyenne empirique des données $(X_{lk})_{l=1,\ldots,n_k}$ du groupe k. On suppose ici que les covariances des différents groupes sont connues et égales entre elles à C, que s_{10} est connu et que $K = 2$ (le résultat restera valable pour $K > 2$ mais nous ne le présentons pas dans ce cas par soucis de clarté). La frontière de séparation entre les deux groupes est donc affine. Nous supposons que seul F_{10} est inconnu, et que l'échantillon d'apprentissage comprend $n_1 = n_0 = n(p)/2$ vecteurs de dimension p. Nous allons proposer une méthode d'estimation du paramètre F_{10} et donner des résultats théoriques lorsque $n(p)$ tend vers l'infini bien moins vite que p.

Nous rappelons que si $q > 0$, la boule $l_p^q(R)$ est l'ensemble des vecteurs $\theta \in \mathbb{R}^p$ tels que

$$\sum_{i=1}^{p} |\theta_i|^q \leq R^q.$$

Nous noterons :

$$\Omega_p(\Theta(R), r) = \left\{ (x, y, C) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathcal{C}_p \ : \ C^{-1/2}(x-y) \in \Theta(R) \text{ et } \|C^{-1/2}(x-y)\|_{\mathbb{R}^p} \geq r \right\} \tag{2.3}$$

où $\mathcal{C}_p$ est l'ensemble des matrices de covariance de rang plein sur $\mathbb{R}^p$. Si $(\mu_0, \mu_1, C) \in \Omega_p(\Theta(R), r)$, nous noterons

$$\mathcal{R}_p(\mu_0, \mu_1, C, \hat{\mathcal{L}}_{10}) = \mathcal{R}(\mathbb{1}_{\hat{V}}), \tag{2.4}$$

où P_0 et P_1 sont les lois gaussiennes sur $\mathbb{R}^p$ de moyennes μ_0 et μ_1 et de covariance C ; $\hat{V}$ est donné par (2.5) et $\mathcal{R}(\mathbb{1}_{\hat{V}})$ par l'équation (1.5).

Procédure. La règle de classification plug-in consiste à affecter une observation X à la classe 1 si X appartient à

$$\hat{V} = \left\{ x \in \mathbb{R}^p \ : \ \hat{\mathcal{L}}_{10}(x) \geq 0 \right\}, \tag{2.5}$$

où

$$\widehat{\mathcal{L}}_{10} = \langle \hat{F}_{10}, X - s_{10} \rangle_{\mathbb{R}^p}.$$

On estime le vecteur directeur $F_{10} = C^{-1} m_{10}$ par $\hat{F}_{10} = C^{-1}\hat{m}_{10}$, où les coefficients de $C^{-1/2}\hat{m}_{10}$ sont donnés par

$$\left(y_{10l} 1_{|y_{10l}| > \lambda_{10}^{FDR}} \right)_{l=1,\dots,p}, \quad \text{où} \ \ y_{10l} = \left(C^{-1/2}(\bar{\mu}_1 - \bar{\mu}_0) \right)_{l=1,\dots,p},$$

et λ_{10}^{FDR} est choisi par la procédure de Benjamini et Hochberg de contrôle du FDR. Nous rappelons que cette procédure (définie et motivée plus en détail Chapitre 4 Partie I) est la suivante. Les $(|y_{10l}|)_l$ sont ordonnés par ordre décroissant :

$$|y_{10(1)}| \geq \cdots \geq |y_{10(p)}| \text{ et } \lambda_{10}^{FDR} = |y_{10(k_{10}^{FDR})}|$$

$$\text{où } k_{10}^{FDR} = \max \left\{ k \in \{1, \dots, p\} \ : \ |y_{10(k)}| \geq \sqrt{\frac{1}{n(p)}} z\left(\frac{b_p k}{2p} \right) \right\},$$

$z(\alpha)$ est le quantile d'ordre α d'une gaussienne et $b_p \in [0, 1/2[$ peut tendre vers 0, mais pas plus vite que $\frac{1}{\log p}$.

Résultat principal.

Theoreme 2.1. *Soit $\hat{V}$ défini par (2.5). Soit $q \in]0, 2[$, $R > 0$. Notons $\eta_p = p^{-\frac{1}{q}} R\sqrt{n(p)}$. Si lorsque p tend vers l'infini, $\eta_p^q \in [\frac{\log^5(p)}{p}, p^{-\delta}]$ pour $\delta > 0$, alors :*

$$\forall r > 0 \quad \sup_{(\mu_0,\mu_1,C) \in \Omega_p(l^q(R), r)} \mathbb{E}_{P^{\otimes n}} \left[\mathcal{R}_p(\mu_0, \mu_1, C, \hat{\mathcal{L}}_{10}) \right] \leq \frac{1 + o_p(1)}{r} \left(\sqrt{2} \frac{\log^{1/2}\left(\frac{p}{R^q n(p)^{q/2}} \right)}{R n^{1/2}(p)} \right)^{\frac{2-q}{2}}, \tag{2.6}$$

où $\mathcal{R}$ est l'erreur d'apprentissage donnée par (2.4), et $P^{\otimes n}$ est la loi de l'échantillon d'apprentissage.

Démonstration. Tout d'abord, d'après le Théorème 1.1 du Chapitre 1 Partie II, on a

$$\mathbb{E}_{P^{\otimes n}}[\mathcal{R}^2(\mathbb{1}_{\hat{V}})] \leq \frac{1}{\|C^{-1/2}(\mu_1 - \mu_0)\|_{\mathbb{R}^p}^2} \mathbb{E}[\|C^{-1/2}(\bar{\mu}_1 - \bar{\mu}_0) - C^{-1/2}(\mu_1 - \mu_0)\|_{\mathbb{R}^p}^2]. \tag{2.7}$$

Puisque s_{10} et $C^{-1/2}$ sont connus, nous pouvons supposer observer $C^{-1/2}(\bar{\mu}_1 - \bar{\mu}_0)$. Il est clair que la matrice de covariance de $C^{-1/2}(\bar{\mu}_1 - \bar{\mu}_0)$ est une matrice diagonale dont les éléments diagonaux sont tous égaux à $\frac{1}{n(p)}$. Ainsi, d'après le Théorème 4.4 (Partie I Chapitre 4 de ce mémoire, en substituant p à m, b_p à q_m, et q à p), on a

$$\sup_{C^{-1/2}(\mu_1-\mu_0)\in l^q(R)} \mathbb{E}\left[\|C^{-1/2}(\bar{\mu}_1 - \bar{\mu}_0) - C^{-1/2}(\mu_1 - \mu_0)\|\right]$$

$$= (1 + o_p(1)) \inf_{\hat{\mu}} \sup_{\mu\in l^q(R)} \mathbb{E}[\|\hat{\mu} - \mu\|^2_{\mathbb{R}^p}],$$

où dans cette dernière équation l'infimum est pris sur tous les estimateurs $\hat{\mu}$ construits à partir de l'observation de $Y \rightsquigarrow \mathcal{N}(\mu, \frac{1}{n})$. D'après le Théorème 5 point $3b.$ de l'article de Donoho et Johnstone [27] (il faut substituer $1/n$ à σ^2, R à r, p à n, q à p et 2 à q dans l'énoncé), si $\eta_p = p^{-1/q} R\sqrt{n}$

$$\inf_{\hat{\mu}} \sup_{\mu\in l^q(R)} \mathbb{E}[\|\hat{\mu} - \mu\|^2_{\mathbb{R}^p}] \sim \left(2\frac{\log(\frac{p}{(n^{1/2}R)^q})}{R^2 n}\right)^{\frac{2-p}{2}}.$$

lorsque $\eta_p \to 0$ (c'est à dire lorsque $p \to \infty$). Ainsi, au vu de (2.7) on a :

$$\forall r > 0, \quad \sup_{(\mu_0,\mu_1,C)\in\Omega_p(l^q(R),r)} \mathbb{E}_{P^{\otimes n}}\left[\mathcal{R}^2_p(\mu_0, \mu_1, C, \hat{\mathcal{L}}_{10})\right] \leq \frac{1 + o_p(1)}{r^2}\left(\sqrt{2}\frac{\log^{1/2}\left(\frac{p}{R^q n(p)^{q/2}}\right)}{Rn^{1/2}(p)}\right)^{2-q},$$

dont le résultat final se déduit par l'inégalité de Jensen :

$$\mathbb{E}_{P^{\otimes n}}\left[\mathcal{R}_p(\mu_0, \mu_1, C, \hat{\mathcal{L}}_{10})\right] \leq \mathbb{E}_{P^{\otimes n}}\left[\mathcal{R}^2_p(\mu_0, \mu_1, C, \hat{\mathcal{L}}_{10})\right]^{1/2}.$$

□

Commentaires. Nous allons faire quelques remarques sur ce théorème.

1. Dans le théorème précédent, on peut réécrire (2.6) en faisant intervenir η_p :

$$\forall r > 0 \quad \sup_{(\mu_0,\mu_1,C)\in\Omega_p(l^q(R),r)} \mathbb{E}_{P^{\otimes n}}\left[\mathcal{R}_p(\mu_0, \mu_1, C, \hat{\mathcal{L}}_{10})\right] \leq \frac{1 + o_p(1)}{r}\left(\sqrt{2}\frac{\log^{1/2}\left(\eta_p^{-q}\right)}{p^{1/q}\eta_p}\right)^{\frac{2-q}{2}},$$

   ceci permet d'analyser la condition sur η_p.
2. La convergence est d'autant plus rapide que q est proche de 0 et d'autant moins rapide que q est proche de 2. En d'autres termes plus $C^{-1/2}(\mu_0 - \mu_1)$ est creux, plus la convergence sera bonne. Nous pensons que la base d'ondelette est une base qui permet une bonne convergence. D'une part elle transforme une vaste famille de courbes en un vecteur creux (propriété de compression) et d'autre par, elle diagonalise quasiment une grande famille d'opérateurs de covariance (propriété de décorrélation).
3. La constante $c(b_p)$ ne dépend pas de $q \in]0, 2[$. Nous pourrions obtenir les mêmes vitesses avec un seuillage universel ($\lambda_U = \frac{1}{n(p)}\sqrt{2\log(p)}$). La constante $\frac{1+o_p(1)}{r^2}$ est dans ce cas bien moins bonne (cf [2]).

4. Il n'existe pas à notre connaissance de résultat donnant une telle convergence (uniformément sur un certain nombre de boules). Par ailleurs, nous ne faisons pas une hypothèse trop forte sur la nature de C. Bickel et Levina [12] ainsi que Fan [32] supposent dans leurs travaux que le rapport entre la plus grande et la plus petite des valeurs propres de C est borné supérieurement indépendamment de la dimension p. Rappelons que si Y est une variable aléatoire gaussienne à valeur dans un espace de Hilbert, alors l'opérateur de covariance de cette variable est nécessairement nucléaire. Ainsi, l'hypothèse faite par les auteurs mentionnés ne permet pas de considérer une mesure gaussienne limite dont le support est un espace de Hilbert.
5. Dans le cas où l'hypothèse $\|C^{-1/2}(\mu_1 - \mu_0)\|_{\mathbb{R}^p} \geq r$ n'est pas vérifiée (les données peuvent tendre à être indistinguable), nous pensons que l'on peut toujours contrôler l'espérance de l'erreur d'apprentissage. En effet, nous pensons que si $\|C^{-1/2}(\mu_1 - \mu_0)\|_{\mathbb{R}^p}$ est petit, l'estimation par seuillage tendra à donner $\|F_{10} - \hat{F}_{10}\|_{L_2(\gamma_C)} \approx \|F_{10}\|_{L_2(\gamma_C)} f(p)$, où $f(p)$ est un terme tendant vers 0. Ceci reste à démontrer.

2.2.2 Cas de covariances et moyennes inconnues : méthode

Dans le reste de ce chapitre, pour $k \in \{1, \ldots, K\}$, $\bar{\mu}_k$ sera la moyenne empirique des observations du groupe k. Nous allons utiliser l'estimateur $\hat{C}_k$ de la matrice de covariance C_k. Ils sera diagonal. Les éléments diagonaux de $\hat{C}_k$ seront notés $(\hat{\sigma}^2_{kq})_{q=1,\ldots,p}$. Pour $q \in \{1, \ldots, p\}$, $k \in \{1, \ldots, K\}$, $\hat{\sigma}^2_{kq}$ sera la version non biaisée de la variance empirique de la coordonnée q des observations $(X_{ikq})_{i=1,\ldots,n_k}$ du groupe k.

Remarque 2.1. *Pour $k \in \{1, \ldots, K\}$, nous choisissons dans toute la suite d'utiliser un estimateurs de C_k qui est diagonal. Ce choix a deux origines. D'une part, il n'est pas possible d'estimer tous les paramètres de C_k sans faire d'approximation (cf Proposition 1.1) et d'autre part, la base d'ondelette diagonalise « quasiment » une large classe d'opérateurs de covariance (cf [55] et [54]).*

Nous noterons

$$\hat{s}_{ij} = (\bar{\mu}_i + \bar{\mu}_j)/2.$$

La règle de classification utilisée consiste à affecter une observation $X \in \mathbb{R}^p$ à la classe k si X appartient à $\hat{V}_k$ donné par (2.1) et

$$\forall (i,j) \in \{1, \ldots, K\}^2, \quad \hat{\mathcal{L}}_{ij} = -\frac{1}{2}\langle \hat{A}_{ij}(x - \hat{s}_{ij}), x - \hat{s}_{ij}\rangle_{\mathbb{R}^p} + \langle \hat{G}_{ij}, x - \hat{s}_{ij}\rangle_{\mathbb{R}^p} - \hat{c}_{ij},$$

où les quantités de cette équation vont être définies par la suite. Pour $(i,j) \in \{1, \ldots, K\}^2$, $i \neq j$, nous allons donner dans l'ordre $\hat{G}_{ij}$ (équation (2.8) ci-dessous), $\hat{A}_{ij}$ (équation 2.10 ci-dessous), et $\hat{c}_{ij}$ (équation 2.11 ci-dessous).

Pour $(i,j) \in \{1, \ldots, K\}^2$, $i \neq j$, on estime le vecteur directeur $G_{ij} = \frac{1}{2}(C_i^{-1} + C_j^{-1})m_{ij}$ par

$$\hat{G}_{ij} = \left(\frac{1}{\sqrt{2}} \left(\frac{1}{\hat{\sigma}^2_{jq}} + \frac{1}{\hat{\sigma}^2_{iq}} \right)^{1/2} y_{ijq} 1_{|y_{ijq}| > \lambda^{FDR}_{ij}} \right)_{q=1,\ldots,p} \tag{2.8}$$

$$\text{où } y_{ijq} = \frac{1}{\sqrt{2}} \left(\frac{1}{\hat{\sigma}^2_{jq}} + \frac{1}{\hat{\sigma}^2_{iq}} \right)^{1/2} (\bar{\mu}_{iq} - \bar{\mu}_{jq}), \tag{2.9}$$

et λ_{ij}^{FDR} est choisi par la procédure de Benjamini et Hochberg de contrôle du FDR. Cette procédure est la suivante. Nous notons $Var_0(y_{ijq})$ la variance de y_{ijq} sous l'hypothèse où $\mu_{iq} = \mu_{jq}$. Le terme

$$\frac{1+\hat{\sigma}_{iq}^2/\hat{\sigma}_{jq}^2}{2n_i} + \frac{1+\hat{\sigma}_{jq}^2/\hat{\sigma}_{iq}^2}{2n_j}$$

est une estimation cette variance sous l'hypothèse où les σ_{jq}^2 sont connus et égals à $\hat{\sigma}_{jq}^2$. Dans la pratique nous substituons ce terme à $Var_0(y_{ijq})$. Les $(|y_{ijq}|/\sqrt{Var_0(y_{ijq})})_{q=1,\dots,p}$ sont ordonnés par ordre décroissant :

$$|y_{ij(1)}|/\sqrt{Var_0(y_{ij(1)})} \geq \cdots \geq |y_{ij(p)}/\sqrt{Var_0(y_{ij(p)})}| \text{ et } \lambda_{ij}^{FDR} = |y_{ij(k_{ij}^{FDR})}|$$

$$\text{où } k_{ij}^{FDR} = \max\left\{k \ : \ |y_{ij(k)}| \geq \sqrt{\frac{1+\hat{\sigma}_{i(k)}^2/\hat{\sigma}_{j(k)}^2}{2n_i} + \frac{1+\hat{\sigma}_{j(k)}^2/\hat{\sigma}_{i(k)}^2}{2n_j}} z\left(\frac{b_p k}{2p}\right)\right\},$$

$z(\alpha)$ est le quantile d'ordre α d'une gaussienne et $b_p \in [0,1[$ tends vers 0 pas plus vite que $\frac{1}{\log p}$.

Dans la pratique, nous avons choisi $b_p = 0.01$, mais il est tout à fait envisageable de réserver une partie de l'échantillon d'apprentissage à l'estimation de b_p. Notons que le choix de b_p n'est pas déterminant pour les propriétés obtenues au Théorème 2.1. Dans la pratique, la différence d'erreur de classification entre les choix $b_p = 0.01$ et $b_p = 0.05$ par exemple, n'est pas significative.

Remarque 2.2. *Cette première partie de la méthode constitue une réduction de dimension. En effet, seules les composantes de $(\hat{G}_{ijq})_{q=1,\dots,p}$ pour lesquelles $|y_{ijq}| \geq \lambda_{ij}^{FDR}$ sont non nulles. L'application linéaire associée à $(\hat{G}_{ijq})_{q=1,\dots,p}$ agit donc seulement dans k_{ij}^{FDR} directions. Notons aussi que pour deux couples $(i,j) \neq (l,m)$, deux estimations de G_{ij} et G_{lm} sont basées sur des réductions de dimension différentes.*

Pour $(i,j) \in \{1,\dots,K\}^2$, $i \neq j$, la matrice A_{ij} est estimée par une matrice diagonale dont les éléments diagonaux sont

$$\hat{a}_{ijq} = \left(\frac{1}{\hat{\sigma}_{iq}^2} - \frac{1}{\hat{\sigma}_{jq}^2}\right) 1_{|w_{ijq}| \geq \eta_{ij}^{FDR}}, \text{ où } \quad w_{ijq} = \hat{\sigma}_{iq}^2 - \hat{\sigma}_{jq}^2, \quad q = 1,\dots,p, \tag{2.10}$$

et le seuil η_{ij}^{FDR} est choisi par le même type de procédure que λ_{ij}^{FDR}. Nous notons $Var_0(w_{ijq})$ la variance de w_{ijq} sous l'hypothèse où $\sigma_{iq} = \sigma_{jq}$. Le terme $\frac{2\hat{\sigma}_{iq}^4}{n_i - 1} + \frac{2\hat{\sigma}_{jq}^4}{n_j - 1}$ en est une estimation que nous utilisons dans la pratique. Les $(|w_{ijq}/\sqrt{Var_0(w_{ijq})}|)_q$ sont ordonnés par ordre décroissant :

$$|w_{ij(1)}/\sqrt{Var_0(w_{ijp})}| \geq \cdots \geq |w_{ij(p)}/\sqrt{Var_0(w_{ijp})}| \text{ et } \eta_{ij}^{FDR} = |w_{ij(k_{ij}^{FDR})}|$$

$$\text{où } k_{ij}^{FDR} = \max\left\{k \ : \ |w_{ij(k)}| \geq \sqrt{\frac{2\hat{\sigma}_{i(k)}^4}{n_i - 1} + \frac{2\hat{\sigma}_{j(k)}^4}{n_j - 1}} z\left(\frac{b_p k}{2p}\right)\right\}.$$

Remarque 2.3. *Cette deuxième partie de la méthode constitue une linéarisation de la règle. En effet, les directions $q \in \{1,\dots,p\}$ dans lesquelles $\hat{a}_{ijq}$ vaut 0 sont des directions dans lesquelles la règle de classification entre les groupes i et j est linéaire. Dans les autres directions, la règle est quadratique.*

L'utilisation de ces méthodes est bien sûr encore motivée par le Théorème 2.1, mais nous n'avons pas pu la justifier théoriquement. Notons que ces dernières années, un grand nombre de travaux se sont tournés vers l'estimation de grandes matrices de covariances ou de leur inverse (voir par exemple [13]).

Pour $(i,j) \in \{1,\dots,K\}^2$, $i \neq j$, on notera enfin :

$$\hat{c}_{ij} = \sum_{q=1}^{p} 1_{|w_{ijq}| \geq \eta_{ij}^{FDR}} \left(\frac{1}{8} \hat{a}_{ijq} (\bar{\mu}_{iq} - \bar{\mu}_{jq})^2 + \frac{1}{2} \log |\det(\hat{\sigma}_{jq}^{-1} \hat{\sigma}_{iq})| \right). \tag{2.11}$$

2.3 Second problème : estimation des densités

Dans le second problème, il s'agit de donner des estimateurs des densités des lois $P_1,\dots,P_K$ des différentes classes. Ceux-ci seront utilisés pour la segmentation d'images hyper-spectrales. Pour $k \in \{1,\dots,K\}$, nous proposons d'estimer la densité f_k de la loi P_k grâce à une procédure de réduction de dimension, et l'estimation d'un sous espace dans lequel les covariances de tous les groupes sont supposées égales. Soient $(e_i)_{i=1,\dots,p}$ les p vecteurs de la base utilisée pour représenter les données. Dans la sous-section suivante, nous allons donner une méthode pour déterminer deux ensembles d'indices : $J_{RD} \subset \{1,\dots,p\}$ (équation (2.15)) et $J_L \subset J_{RD}$ (équations (2.17)). Sur le sous espace $Vect((e_i)_{i \notin J_{RD}})$ les moyennes et covariances de groupes seront supposées égales et sur le sous espace $E_L^k = Vect((e_i)_{i \in J_{RD} \setminus J_L})$ les covariances de groupe (seulement) seront supposées égales.

Nous notons $\bar{\mu}_k$ la moyenne empirique des observations du groupe k : $(X_{ik})_{i=1,\dots,n_k}$, $\bar{\mu}$ la moyenne empirique de toutes les observations : $(X_{ik})_{i=1,\dots,n_k;k=1,\dots,K}$, $\hat{C}_k$ la matrice de covariance empirique des observations du groupe k, et $\hat{C}$ la matrice de covariance empirique de toutes les observations. Pour $k \in \{1,\dots,K\}$ fixé, l'estimateur $\hat{f}_k$ de f_k est une densité gaussienne. On le construit donc à partir de sa moyenne Moy_k et de sa covariance Cov_k. L'estimateur de la moyenne du groupe k est alors donné par le vecteur de $\mathbb{R}^p$ Moy_k définit par :

$$\forall q \in \{1,\dots,p\}, \quad Moy_{kq} = \quad \bar{\mu}_{kq} \text{ si } q \in J_{RD} \text{ et } \bar{\mu}_q \text{ sinon.}$$

L'estimateur de la covariance du groupe k est donné par la matrice diagonale Cov_k d'éléments diagonaux définis par :

$$\forall q \in \{1,\dots,p\}, \quad Cov_{kqq} = \hat{C}_{kqq} \text{ si } q \in J_L \cap J_{RD} \text{ et } \hat{C}_{qq} \text{ sinon.}$$

Finalement, l'estimateur de f_k est donné par

$$\hat{f}_k(x) = \frac{1}{\sqrt{(2\pi)^p Det(Cov_k)}} e^{-\frac{1}{2}\langle Cov_k^- (x - Moy_k), x - Moy_k \rangle_{\mathbb{R}^p}} \tag{2.12}$$

(où si A est une matrice symétrique semi-définie positive, A^- est l'inverse généralité de moore penrose associée (voir Chapitre 1 Partie II Proposition 1.1)). Nous voulons que ces estimateurs soient construits pour qu'une règle de décision définie par (2.1) soit efficace.

2.3.1 Réduction de dimension

La première sélection consiste en une sélection des directions sur lesquelles la règle agit : c'est la réduction de dimension. Nous introduisons la variable suivante mesurant la pertinence d'une direction q, pour la classification :

$$\forall q \in \{1,\dots,p\},\ \ \mathcal{I}_{RD}(q) = \sum_{i<j}\sum_{m=1}^{n_i}\sum_{l=1}^{n_j}\frac{1}{2}\left(\frac{1}{\bar{\sigma}_{jq}}+\frac{1}{\bar{\sigma}_{iq}}\right)\frac{1}{n_i n_j}(X_{miq}-X_{ljq})^2.$$

la première somme étant prise sur tous les couples $(i,j) \in \{1,\dots,K\}^2$ tels que $i<j$ et $\bar{\sigma}_{jq}$ est un estimateur robuste de σ_{jq} défini par

$$\bar{\sigma}_{kq} = MAD((X_{ikq})_{i=1,\dots,n_k})/0.6745, \tag{2.13}$$

où MAD est la médiane des écarts, en valeur absolue, à la médiane. Nous en déduisons une variable aléatoire S_q^{RD} centrée réduite sous l'hypothèse que dans cette direction q, y_{ijq} (défini par (2.9)) est d'espérance nulle :

$$\forall q \in \{1,\dots,p\},\ \ S_q^{RD} = \frac{\mathcal{I}_{RD}(q) - \mathbb{E}_0\left[\mathcal{I}_{RD}(q)\right]}{\sqrt{\mathrm{Var}_0(\mathcal{I}_{RD}(q))}}. \tag{2.14}$$

Nous choisissons J_{RD} comme étant l'ensemble des directions q pour lesquelles S_q^{RD} est supérieur à un seuil β^{FDR}. Ce seuil est encore une fois obtenu par la méthode de Benjamini et Hochberg associée aux données $S_1^{RD},\dots,S_p^{RD}$. L'ensemble d'indices finalement obtenu est

$$J_{RD} = FDR(S_1^{RD},\dots,S_p^{RD},0.01,1), \tag{2.15}$$

où la procédure $FDR(X,q,\sigma)$ est définie Partie I Chapitre 4.

La procédure de sélection correspondante a une interprétation statistique intéressante. Elle correspond à tester simultanément pour $q=1,\dots,p$ les hypothèses

- H_{0q} : la proportion de variance inter-groupe dans la variance globale sur la direction ψ_q est nulle,
  contre
- H_{1q} : la proportion de variance inter-groupe dans la variance globale sur la direction ψ_q est non nulle.

Les sous-espaces de petite dimension ayant la plus forte variabilité inter-groupe sont ceux dans lesquels les données sont le mieux séparées (i.e le plus facilement classifiables), cette procédure paraît donc naturelle. L'intérêt d'une direction dans laquelle la variabilité inter-groupe est forte est illustré par la Figure 1.1 du Chapitre 1. Puisque à chacune des hypothèses qui modélisent notre problème correspond une analyse de la variance, on peut parler d'analyse de la variance multiple. Notons que nous aurions pu essayer d'utiliser la loi du χ^2 pour effectuer ce test multiple.

La proposition suivante donne $\mathrm{Var}_0(\mathcal{I}_{RD}(q))$ et $\mathbb{E}_0\left[\mathcal{I}_{RD}(q)\right]$ en substituant σ_{iq} à $\bar{\sigma}_{iq}$. Dans la pratique, ces quantités sont estimées par plug-in via $\bar{\sigma}_{iq}$.

Proposition 2.1. *Si l'on note, pour* $q \in \{1,\dots,p\}$

$$T_q = \sum_{i<j}\sum_{m=1}^{n_i}\sum_{l=1}^{n_j}\alpha_{ij}(X_{miq}-X_{ljq})^2, \quad et \ \ \alpha_{ijq} = \frac{1}{2}\left(\frac{1}{\sigma_{jq}}+\frac{1}{\sigma_{iq}}\right)\frac{1}{n_i n_j},$$

on a alors :

$$\mathbb{E}_0\left[T_q\right] = \sum_{i=1}^{K} \sum_{j \neq i} \alpha_{ijq} n_i n_j \sigma_{iq}^2,$$

$$\mathrm{Var}_0(T_q) = 4 \sum_{i=1}^{K} \sum_{j \neq i} \sum_{k \neq i} \alpha_{ik} \alpha_{ijq} n_i n_j n_k \sigma_{iq}^4.$$

La démonstration de ce résultat est donnée par le Lemme 3.6 du chapitre suivant.

2.3.2 Linéarisation de la règle

Nous introduisons la variable aléatoire suivante mesurant la pertinence dans une direction q de l'aspect quadratique de la règle utilisée :

$$\forall q \in \{1, \ldots, p\}, \quad \mathcal{I}_L(q) = \sum_{i<j} \left(\hat{\sigma}_{iq}^2 - \hat{\sigma}_{jq}^2\right)^2.$$

Nous en déduisons une variable aléatoire S_q^L centrée réduite sous l'hypothèse que dans cette direction q, $\sigma_i = \sigma_j$:

$$\forall q \in \{1, \ldots, p\}, \quad S_q^L = \frac{\mathcal{I}_L(q) - \mathbb{E}_0\left[\mathcal{I}_L(q)\right]}{\sqrt{\mathrm{Var}_0(\mathcal{I}_L(q))}}. \tag{2.16}$$

Nous choisissons J_L comme étant l'ensemble des directions q pour lesquelles S_q^L est supérieur à un seuil β^{FDR}. Ce seuil est encore une fois obtenu par la méthode de Benjamini et Hochberg associée aux données $(S_q^L)_{q \in J}$. L'ensemble d'indices finalement obtenu est

$$J_L = FDR((S_q^L)_{q \in J_{RD}}), 0.01, 1), \tag{2.17}$$

où la procédure $FDR(X, q, \sigma)$ est définie Partie I Chapitre 4.

La procédure d'estimation correspondante a une interprétation statistique intéressante. Elle correspond à tester simultanément pour $q \in J_{RD}$ les hypothèses

- H_{0q} : la variabilité des variances de groupe est nulle dans la direction q,
  contre
- H_{1q} : la variabilité des variances de groupe est non nulle dans la direction q.

Parmi les sous espaces de petite dimension ayant la plus forte variabilité intergroupe, ceux dans lesquels les variances de groupe varient d'un groupe à l'autre, sont mieux séparés par une règle quadratique que par une règle linéaire. Cette procédure paraît donc naturelle. L'intérêt d'une direction dans lesquelles les variances de groupe varient d'un groupe à l'autre est illustré par la Figure 1.2 du Chapitre 1. Notons que ici aussi, au lieu d'utiliser une statistique centrée réduite (ayant pour vocation d'être gaussienne), nous aurions pu utiliser la loi de Fisher pour construire nos tests.

La proposition suivante donne $\mathrm{Var}_0(\mathcal{I}_L(q))$ et $\mathbb{E}_0\left[\mathcal{I}_L(q)\right]$ que l'on estime dans la pratique par plug-in.

Proposition 2.2. *Avec les notations précédentes, on a :*

$$\forall q \in \{1,\dots,p\}, \quad \mathbb{E}_0\left[\mathcal{I}_L(q)\right] = \sum_{i=1}^{K}\sum_{j\neq i}\frac{\sigma_{jq}^4(n_i-3)}{n_i-1} - \sigma_{iq}^2\sigma_{jq}^2.$$

et

$$\mathrm{Var}_0\left(\mathcal{I}_L(q)\right)\left(1+\sum_{k=1}^{K} o_{n_k}(1)\right) = 4\sum_{i<j}\left(\frac{\sigma_{iq}^8}{n_i}+\frac{\sigma_{iq}^8}{n_i}+\frac{(n_i+n_j)\sigma_{iq}^4\sigma_{jq}^4}{n_i n_j}\right).$$

2.4 Application aux données médicales et étude de l'efficacité de notre méthode

Nous allons étudier l'efficacité de la première procédure dans le problème de classification. Pour cela nous comparons notre méthode à celle utilisée par Rossi et Villa [63] sur les données de la base TIMIT. Pour finir, nous testons les performances de notre procédure sur des données médicales.

2.4.1 Méthodes utilisées par Rossi et Villa, classification à deux classes

Rossi et Villa utilisent une adaptation de machines à support vectoriel (SVM) avec différents types de noyaux. Rappelons que la procédure SVM correspond à construire une fonction frontière f affine donnée par

$$f(x) = \langle w, x\rangle_{\mathbb{R}^p} + b,$$

où w et b sont solution d'un problème d'optimisation du type :

$$\min_{w,b,\xi} \|w\|_{\mathbb{R}^p}^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{sous } y_i\left(\langle w, x_i\rangle_{\mathbb{R}^p} + b\right) \geq 1-\xi_i, \quad \xi_i \geq 0 \quad i=1,\dots,n$$

où $(x_i, y_i)_{i=1,\dots,n}$ sont les couples (observations, labels) de l'échantillon d'apprentissage.

La base de donnée TIMIT a notamment été étudiée par Hastie et Al. dans [41]. Elle comprend les phonèmes « aa » et « ao » pronocés par un grand nombre de personnes différentes. Les enregistrements correspondant sont assimilables à des courbes observées à une fréquence d'échantillonnage assez fine. Plus précisément, une courbe peut être assimilée à un vecteur de dimension $p = 256$. L'ensemble d'apprentissage est composé de 519 « aa » et 759 « ao » et l'ensemble de test est composé de 176 « aa » et 263 « ao » . Ainsi, les courbes $(x_i)_{i=1,\dots,519}$ sont celles qui correspondent à la prononciation du phonème « aa » et le label $y_i = 0$ leur est associé. Le label « 1 » est associé aux autres courbes qui correspondent à la prononciation du phonème « ao » . La Méthode de Rossi et Villa et la notre donnent un taux d'erreur de classification du même ordre : 20% d'erreur.

2.4.2 Application aux données médicales

Le problème médical est le suivant. Des spectres sélectionnés par les médecins comme représentatifs de certaines tumeurs nous ont été fournis. Il a été difficile d'obtenir un grand nombre de spectres pour chaque type de tumeur cérébrale. Nous avons retenu cinq groupes de spectres correspondant à cinq groupes de tissus : les glioblastomes de type A[1], les glioblastomes de type B, les Meningiomes, les Metastases et les tissus sains. La base fournie par les médecins contient 21 glioblastomes de type A, 9 glioblastomes de type B, 16 Méningiomes, 18 métastases et 9 tissus sains, c'est-à-dire en tout, 75 spectres discrétisés en 1024 bandes spectrales. Nous donnons le tracé des spectres des groupes considérés à la Figure 2.1. Afin de tester notre procédure, nous avons utilisé une stratégie de type « Leave one out » [2]. La Figure 2.3 nous permet de confirmer expérimentalement le caractère quasiment optimal de la dimension choisie dans le cas de deux classes.

Nous avons testé différentes configurations résumées dans le tableau de la Figure 2.2. En définitive, les erreurs de classification restent assez importantes (rappelons tout de même que dans le cas de 4 classes ayant les mêmes fréquences d'apparition, une règle qui ferait un choix au hasard aurait un taux d'erreur de 75%). Nous donnons à cela deux origines extérieures à l'algorithme utilisé.

La physique théorique prévoit qu'un spectre associé à une tumeur donnée, par exemple un Glioblastome, est une variable aléatoire $Y \in \mathbb{R}^p$ qui a une assez faible variabilité. Ainsi, on devrait pouvoir aisément séparer les spectres associés à des groupes différents. Malheureusement, des problèmes pratiques liés à l'instrumentation, et sur lesquels les physiciens travaillent activement, font que l'on observe un spectre $Z \in \mathbb{C}^p$ (à valeurs complexes) pour lequel il existe une série d'angles $(\psi_q)_{q=1\ldots,p}$ tels que

$$\forall q \in \{1, \ldots, p\} \quad Y_q(w) = \Re(e^{i\psi_q(w)} Z_q(w)).$$

Cette série d'angle n'est pas connu. La physique théorique de l'instrumentation indique qu'il existe un couple de réels (a, b) tels que

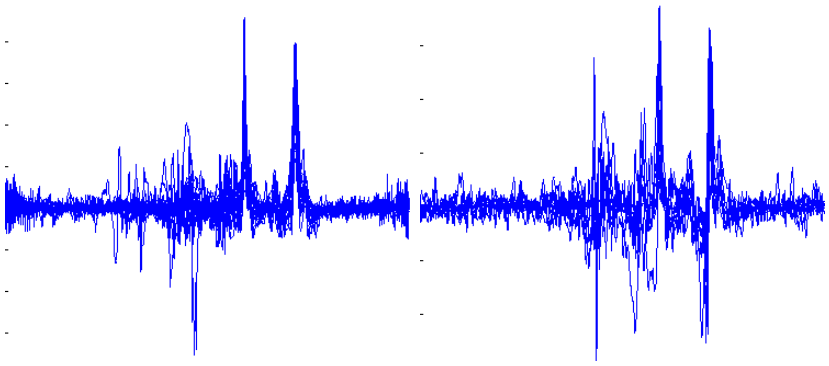
$$\forall q \in \{1, \ldots, p\} \;\; \psi_q(w) = a(w)q + b(w).$$

Les méthodes pour obtenir a et b ne sont pas encore suffisament au point. Nous avons choisi de demander aux médecins d'appliquer aux données un rephasage afin que les parties réelles des spectres soient assez homogènes au sein d'un groupe donné, et nous n'avons gardé que les parties réelles des spectres. Le rephasage effectué par les médecins n'est pas optimal et le déphasage résiduel est à l'origine d'une forte disparité des spectres observés au sein de chaque groupe. Cette disparité est visible sur la Figure 2.1. L'incorporation du phénomène de déphasage dans un algorithme de classification, ainsi que la prise en compte de la nature complexe des données fera l'objet d'un travail ultérieur. Notons juste que ce qui est un problème de déphasage dans le domaine de Fourier est un problème de recalage dans le domaine temporel.

Enfin, les échantillons d'apprentissage sont encore de taille trop petite. Nous avons bon espoir de voir ces tailles grandir dans les années à venir.

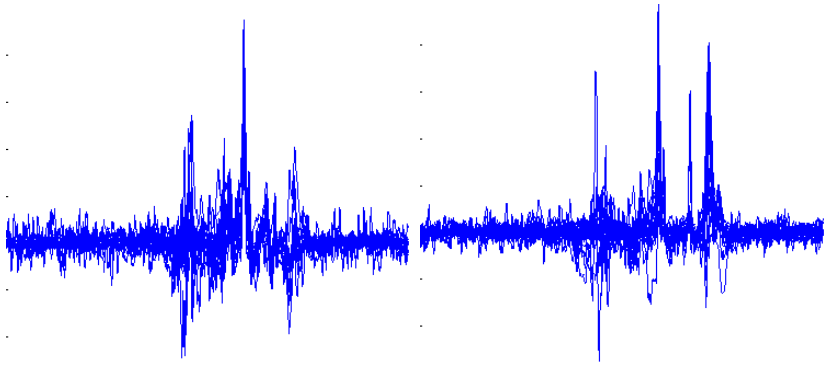
[1] Le groupe des glioblastomes a une trop grande variabilité, aussi nous avons choisis de le scinder en deux groupes. La séparation choisie a un lien avec la gravité du Glioblastome : le type B est plus grave que le type A

[2] Si l'on dispose en tout de n données pour l'apprentissage et le test confondu, la stratégie leave one out consiste pour $i = 1, \ldots, n$ à utiliser les données $\{1, \ldots, n\} \setminus \{i\}$ pour apprendre et la donnée $\{i\}$ pour tester. Ainsi, l'échantillon d'apprentissage est toujours de taille $n - 1$ et virtuellement l'échantillon de test est de taille n.



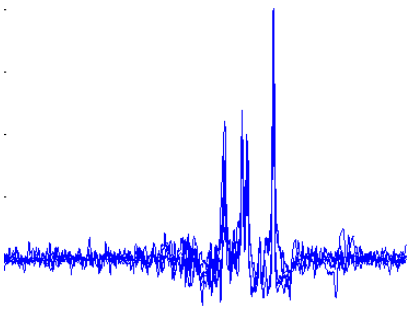
(a) 21 glioblastomes A

(b) 9 glioblastomes B



(c) 16 Meningiomes

(d) 18 métastases



(e) 9 tissus sains

FIG. 2.1 – Spectres de l'échantillon d'apprentissage

Groupes présents	tous	tous sauf Métastases	Glioblastomes type A et Méningiomes
Taux d'erreur	43 %	30 %	5%

FIG. 2.2 – Configurations retenue et taux d'erreur de classification dans chaque cas

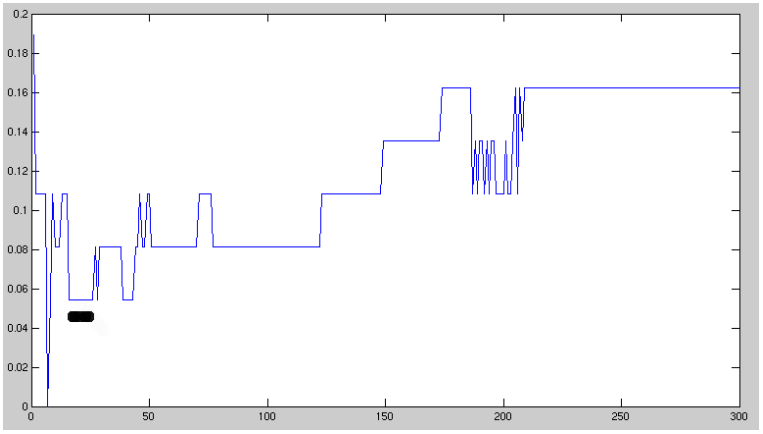


FIG. 2.3 – Taux d'erreur de classification (problème à deux groupes : Méningiomes / Glioblastome A) en fonction de la dimension sélectionnée. La dimension sélectionnée par notre algorithme est dans la zone marquée de points noirs.

Chapitre 3

Démonstration des résultats

> The blessings of dimensionality are less widely noted, but they include the concentration of measure phenomenon (so-called in the geometry of Banach spaces), which means that certain random fluctuations are very well controlled in high dimensions and the success of asymptotic methods, used widely in mathematical statistics and statistical physics, which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated.
>
> Donoho

Ce chapitre est consacré à la démonstration des résultats énoncés dans les Chapitres 1 et 2. Il est divisé en trois sections. Dans la première nous démontrons les résultats relatifs à la procédure LDA. Dans la seconde, nous démontrons les résultats relatifs à la procédure QDA. Dans la troisième, nous donnons la preuve des résultats techniques restant.

Soit $\mathcal{X}$ un espace de Banach (la grande majeur parti de ce qui est dit par la suite concerne $\mathcal{X} = \mathbb{R}^p$), muni de sa tribu Borélienne et d'une mesure gaussienne γ. Dans tout ce chapitre, si f est une application mesurable, nous noterons :

$$V_f = \{x \in \mathcal{X} \quad : \quad f(x) \geq 0\}. \tag{3.1}$$

3.1 Cas de la procédure LDA en dimension finie

Dans toute cette section, $\mathcal{X} = \mathbb{R}^p$. Rappelons que γ_C est la mesure gaussienne de covariance C, et $\gamma_{C,\mu}$ est la mesure gaussienne de covariance C et de moyenne μ. Si p est un entier positif, nous noterons $\gamma_p = \gamma_{I_p,0}$. Rappelons que

$$F_{10} = C^{-1}m_{10}, \quad m_{10} = \mu_1 - \mu_0, \quad s_{10} = \frac{\mu_1 + \mu_0}{2},$$

où μ_1, μ_0 et C sont les moyennes et la covariance (commune) des lois $P_1 = \gamma_{C,\mu_1}$ et $P_0 = \gamma_{C,\mu_0}$ des groupes 1 et 0. Avec la notation définie par l'équation (3.1), les règles optimales et plug-in (définies par les équations (1.2) et (1.7)) peuvent être réécrites de la manière suivante. On affecte une nouvelle donnée X à la classe 1 si X appartient à $V = V_{\langle F_{10}, x - s_{10}\rangle_{\mathbb{R}^p}}$ dans le cas de la règle optimale et si X appartient à $\hat{V} = V_{\langle \hat{F}_{10}, x - \hat{s}_{10}\rangle_{\mathbb{R}^p}}$ dans le cas de la règle plug-in. Ainsi, l'erreur d'apprentissage définie Chapitre 1 Partie I est donnée par

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) = \frac{1}{2}\left(\gamma_{C,\mu_0}\left(X \in \hat{V} \setminus V\right) + \gamma_{C,\mu_1}\left(X \in V \setminus \hat{V}\right)\right). \tag{3.2}$$

Nous noterons γ_1 la mesure gaussienne centrée réduite sur $\mathbb{R}$, et ferons un large usage du fait que $\gamma_1([0;u]) \leq \frac{u}{\sqrt{2\pi}}$. Si γ est une mesure sur $\mathbb{R}^p$, $\|\Pi_x^{\perp} e\|_{L_2(\gamma)}$ sera la norme de la projection orthogonale dans $L_2(\gamma)$ du vecteur $e \in L_2(\gamma)$ sur l'hyperplan orthogonal à $x \in L_2(\gamma)$. Pour finir, rappelons que α (défini par (1.8) est l'angle dans $L_2(\gamma_C)$ entre F_{10} et $\hat{F}_{10}$. Cette quantité jouera un rôle très important dans toute cette section. Afin d'abréger les notations, nous noterons dans ce chapitre $\mathcal{R}$ au lieu de $\mathcal{R}(\mathbb{1}_{\hat{V}})$.

Le théorème qui va suivre exprime l'erreur d'apprentissage $\mathcal{R}$ entre autre en fonction de α. Dans sa démonstration, nous donnons une autre formulation de l'angle α dans la géométrie euclidienne de $\mathbb{R}^p$.

Theoreme 3.1. *Soit $d_0 = \langle \hat{F}_{10}, \hat{s}_{10} - s_{10}\rangle_{\mathbb{R}^p}$. L'erreur d'apprentissage $\mathcal{R}$ donnée par (3.2) est une fonction notamment de α (donné par (1.8)) qui vérifie :*

$$\forall \alpha \in [-\pi, \pi] \quad \mathcal{R}(\alpha) = \mathcal{R}(-\alpha).$$

Cette erreur vérifie les inégalités suivantes.

Si $\alpha \geq \frac{\pi}{2}$, alors $\mathcal{R} \geq \frac{1}{2}$.

Si $0 \leq \alpha < \frac{\pi}{2}$, on a $\mathcal{R} \leq \frac{1}{2}$ et on distingue plusieurs cas.

1. *Si $|d_0| \leq \frac{1}{4}|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|$, on a :*

$$e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}} \frac{1}{4}\left(\frac{\alpha}{2\pi} + \frac{1}{2}\gamma_1\left(\left[0; \frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^{\perp}} \hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right)\right) \leq \mathcal{R}, \tag{3.3}$$

et

$$\mathcal{R} \leq e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}\cos(\alpha)^2}{32}}\left(\frac{\alpha}{2\pi} + \gamma_1\left(\left[0; (1+\tan(\alpha))\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^{\perp}} \hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right)\right). \tag{3.4}$$

2. *Si $\frac{1}{4}|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}| < |d_0| \leq \frac{1}{2}|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|$, on a :*

$$e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{2}} \frac{1}{4}\left(\frac{1}{2}\gamma_1\left(\left[0; \frac{\|F_{10}\|_{L_2(\gamma_C)}}{4}\right]\right) + \frac{\alpha}{2\pi}\right) \leq \mathcal{R} \tag{3.5}$$

$$\mathcal{R} \leq \frac{\alpha}{2\pi} + \gamma_1\left(\left[0; (1+\tan(\alpha))\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^{\perp}} \hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right). \tag{3.6}$$

3. Si $\frac{1}{2}|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}| < |d_0|$*, on a :*

$$\frac{\alpha}{4\pi} + \frac{1}{4}\gamma_1\left(\left[0; \frac{\|F_{10}\|_{L_2(\gamma_C)}}{2}\right]\right) \leq \mathcal{R} \leq \frac{\alpha}{2\pi} + \gamma_1\left(\left[0; (1+\tan(\alpha))\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^{\perp}}\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right). \tag{3.7}$$

4. Si $|d_0| = 0$*, alors on a :*

$$e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}}\frac{\alpha}{2\pi} \leq \mathcal{R}. \tag{3.8}$$

Démonstration. Pour commencer, notons que (3.2) implique

$$\mathcal{R} = \frac{1}{2}\left(\gamma_{C,s_{10}}\left(X \in \left(\hat{V} \setminus V - \frac{m_{10}}{2}\right)\right) + \gamma_{C,s_{10}}\left(X \in \left(V \setminus \hat{V} + \frac{m_{10}}{2}\right)\right)\right). \tag{3.9}$$

La démonstration se décompose alors de la manière suivante : nous ramenons le problème à un problème dans lequel $C = I_p$, puis nous réduisons le problème de $\mathbb{R}^p$ correspondant à un problème dans $\mathbb{R}^2$. Le Théorème 3.1 repose alors sur le Lemme 3.1 démontré à la section suivante. Dans $\mathbb{R}^2$ la démonstration de ce Lemme est essentiellement géométrique, mais il est possible d'avoir une vision du problème directement dans un sous espace de $L_2(\gamma_C)$ de dimension 2 bien choisi et de traiter le problème de manière géométrique avec la géométrie induite par la mesure γ_C. Nous avons choisi de décomposer les étapes de notre raisonnement pour finalement démontrer les équations du théorème dans la géométrie euclidienne de $\mathbb{R}^2$ (celle induite par γ_2 la mesure gaussienne centrée réduite dans $\mathbb{R}^2$). Ainsi la partie intéressante de cette preuve est toute contenue dans le Lemme 3.1, et ce qui précède ce lemme n'est que la réécriture du problème successivement dans les espaces

$$L_2(\gamma_{C,s_{10}}) \to L_2(\gamma_C) \to L_2(\gamma_p) \to L_2(\gamma_2) \simeq \mathbb{R}^2.$$

Pour se ramener au cas centré réduit dans $\mathbb{R}^p$. Nous allons nous ramener au cas gaussien centré réduit en dimension p, en montrant que

$$\mathcal{R} = \frac{1}{2}\gamma_p\left((V_{\langle .,G_p\rangle_{\mathbb{R}^p}} \setminus V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}) - \frac{G_p}{2}\right) + \frac{1}{2}\gamma_p\left((V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0} \setminus V_{\langle .,G_p\rangle_{\mathbb{R}^p}}) + \frac{G_p}{2}\right), \tag{3.10}$$

où γ_p est la mesure gaussienne centrée réduite sur $\mathbb{R}^p$, $d_0 = \langle \hat{F}_{10}; \hat{s}_{10} - s_{10}\rangle_{\mathbb{R}^p}$,

$$G_p = C^{1/2}F_{10} = C^{-1/2}m_{10}, \;\; \hat{G}_p = C^{1/2}\hat{F}_{10} \;\text{ et }\; e_p = C^{1/2}(\hat{F}_{10} - F_{10}). \tag{3.11}$$

Pour démontrer (3.10), il suffit d'appliquer successivement les deux propriétés suivantes :

1. Si A est une partie de $\mathbb{R}^p$, alors $\gamma_{C,s_{10}}(A) = \gamma_p(C_p^{-1/2}(A - s_{10}))$.
2. Si $A = \{x \in \mathbb{R}^p \;:\; \langle x - s_{10}, v\rangle \geq b\}$ alors $C_p^{-1/2}(A - s_{10}) = \{x \in \mathbb{R}^p \;:\; \langle x, C^{1/2}v\rangle \geq b\}$.

La première propriété nous permet d'obtenir

$$\mathcal{R} = \frac{1}{2}\gamma_p\left(C^{-1/2}\left(V_{\langle .-s_{10},F_{10}\rangle_{\mathbb{R}^p}} \setminus V_{\langle .-s_{10},F_{10}+\hat{F}_{10}-F_{10}\rangle_{\mathbb{R}^p}+d_0} - s_{10}\right) - \frac{G_p}{2}\right)$$

$$+\frac{1}{2}\gamma_p\left(C^{-1/2}\left(V_{\langle .-s_{10},F_{10}+\hat{F}_{10}-F_{10}\rangle_{\mathbb{R}^p}+d_0}\setminus V_{\langle .-s_{10},F_{10}\rangle_{\mathbb{R}^p}}-s_p\right)+\frac{G_p}{2}\right),$$

et la deuxième implique alors au vu de (3.9) l'égalité (3.10).

Gardons en mémoire pour la suite que les quantités introduites (équation (3.11)) par le changement de géométrie effectué vérifient :

$$\|G_p\|_{\mathbb{R}^p}=\|F_{10}\|_{L_2(\gamma)},\ \ \|\hat{G}_p\|_{\mathbb{R}^p}=\|\hat{F}_{10}\|_{L_2(\gamma)},\ \ \|e_p\|_p=\|F_{10}-\hat{F}_{10}\|_{L_2(\gamma_C)}, \tag{3.12}$$

et α (défini par l'équation (1.8)) est l'angle dans $\mathbb{R}^p$ entre G_p et $\hat{G}_p$.

Le problème se ramène au cas de la dimension 2. Nous allons maintenant démontrer l'égalité :

$$\mathcal{R}=\frac{1}{2}\gamma_2\left(Q_-^a-y^+\right)+\frac{1}{2}\gamma_2\left(Q_-^b-y^-\right), \tag{3.13}$$

où γ_2 est la mesure gaussienne centrée réduite dans $\mathbb{R}^2$, Q_-^a, Q_-^b, y_+ et y_- vont être définis dans la suite. Notons seulement que Q_-^a et Q_-^b sont deux parties de $\mathbb{R}^2$, y_+ et y_- sont deux vecteurs de $\mathbb{R}^2$ et toutes ces quantités sont illustrées par la Figure 3.1. Dans la suite, nous allons noter $\tilde{e}_p=\Pi_{G_p^\perp}e_p$ la projection orthogonale de e_p sur l'orthogonal de G_p dans $\mathbb{R}^p$. Nous supposerons que $\|\tilde{e}_p\|_{\mathbb{R}^p}\neq 0$, la partie du résultat concernant le cas $\|\tilde{e}_p\|_{\mathbb{R}^p}=0$ s'obtenant de manière directe. Le calcul de $\mathcal{R}$ se ramène à un calcul dans le sous espace de dimension deux, M_p, engendré par les deux vecteurs G_p et $\tilde{e}_p$. Pour comprendre l'origine de cette affirmation, notons

$$M_p^\perp=\{x\in\mathbb{R}^p\ \ tq\ \ \forall u\in M_p\ :\ \ \langle x,u\rangle_{\mathbb{R}^p}=0\}$$

l'orthogonale de M_p dans $\mathbb{R}^p$ et remarquons que pour tout $z_1\in M_p$ $z_2\in M_p^\perp$ on a :

$$V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}\setminus V_{\langle .,G_p\rangle_{\mathbb{R}^p}}+z_1+z_2=V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}\setminus V_{\langle .,G_p\rangle_{\mathbb{R}^p}}+z_1$$

et

$$V_{\langle .,G_p\rangle_{\mathbb{R}^p}}\setminus V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}+z_1+z_2=V_{\langle .,G_p\rangle_{\mathbb{R}^p}}\setminus V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}+z_1.$$

On peut donc affirmer par les propriétés tensorielles de γ_p et l'équation (3.10), que on a :

$$\begin{aligned}\mathcal{R}\ \ &=\ \ \frac{1}{2}\gamma_2\left(M_p\cap(V_{\langle .\,,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}\setminus V_{\langle .\,,G_p\rangle_{\mathbb{R}^p}}-\frac{G_p}{2})\right) &(3.14)\\ &\quad+\frac{1}{2}\gamma_2\left(M_p\cap(V_{\langle .\,,G_p\rangle_{\mathbb{R}^p}}\setminus V_{\langle .\,,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}+\frac{G_p}{2})\right). &(3.15)\end{aligned}$$

Aussi, dans toute la suite nous assimilerons M_p à $\mathbb{R}^2$, D et $\hat{D}$ seront les droites de M_p d'équations $\langle .,G_p\rangle_{\mathbb{R}^p}=0$ et $\langle .,G_p+e_p\rangle_{\mathbb{R}^p}+d_0=0$. Un calcul simple permet de montrer que ces droites se coupent en a_p défini par

$$a_p=-d_0\frac{\tilde{e}_p}{\|\tilde{e}_p\|^2_{\mathbb{R}^p}}. \tag{3.16}$$

Ainsi

$$V_{\langle .\,,G_p\rangle_{\mathbb{R}^p}}=V_{\langle .-a_p,G_p\rangle_{\mathbb{R}^p}}\ \ \text{et}\ \ V_{\langle .\,,G_p+e_p\rangle_{\mathbb{R}^p}+d_0}=V_{\langle .-a_p,G_p+e_p\rangle_{\mathbb{R}^p}},$$

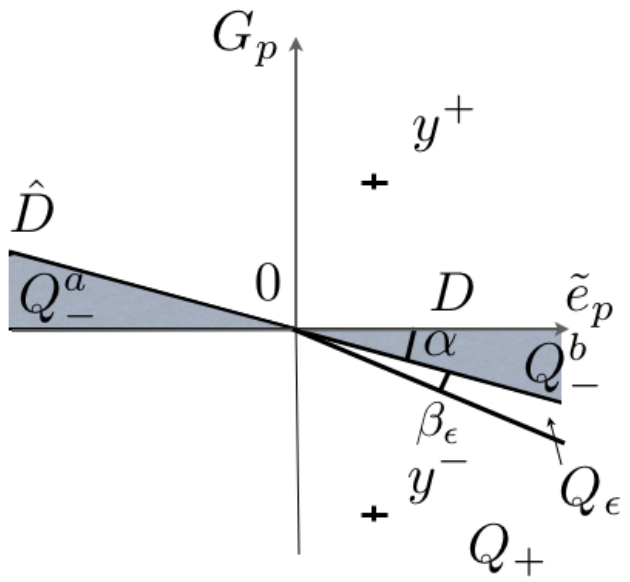


FIG. 3.1 – Figure définissant Q_-^a, Q_-^b, Q_+, et Q_ϵ pour le Lemme 3.1

et avec les mêmes calculs que ceux utilisés pour obtenir l'équation (3.10), l'équation (3.14) devient :

$$\mathcal{R} = \frac{1}{2}\gamma_2\left(M_p \cap (V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}} \setminus V_{\langle .,G_p\rangle_{\mathbb{R}^p}}) - \frac{G_p}{2} + a_p\right) \tag{3.17}$$

$$+\frac{1}{2}\gamma_2\left(M_p \cap (V_{\langle .,G_p\rangle_{\mathbb{R}^p}} \setminus V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}}) + \frac{G_p}{2} + a_p\right). \tag{3.18}$$

Notons que pour des raisons de symétrie, on peut supposer $d_0 \geq 0$ sans restriction de généralité. C'est ce que nous ferons. Dans la suite, nous allons noter

$$y^+ = \frac{G_p}{2} - a_p \text{ et } y^- = -\frac{G_p}{2} - a_p, \tag{3.19}$$

les coordonnées de y^+ dans le repère orthonormé obtenu à partir du repère orthogonal $(0, \tilde{e}_p, G_p)$ seront notés (y_h, y_v) et valent donc $(\frac{d_0}{\|\tilde{e}_p\|_{\mathbb{R}^p}}, \frac{\|G_p\|_{\mathbb{R}^p}}{2})$. Nous noterons aussi

$$Q_-^a = M_p \cap (V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}} \setminus V_{\langle .,G_p\rangle_{\mathbb{R}^p}}) \text{ et } Q_-^b = M_p \cap (V_{\langle .,G_p\rangle_{\mathbb{R}^p}} \setminus V_{\langle .,G_p+e_p\rangle_{\mathbb{R}^p}}). \tag{3.20}$$

Toutes ces notations sont illustrées à la Figure 3.1, et on a finalement l'équation (3.13). On remarque avec cette Figure et cette équation que si l'on change α en $-\alpha$, $\mathcal{R}$ reste inchangé, que si $0 < \alpha \leq \pi/2$ alors $\mathcal{R} \leq \frac{1}{2}$ et que si $\pi \geq \alpha \geq \pi/2$ alors $\mathcal{R}_p \geq 1/2$. Ainsi, nous supposerons maintenant que $\alpha \in [0, \pi/2]$. La suite de la démonstration du théorème repose sur le lemme suivant.

Lemme 3.1. *Soient, Q_+ et Q_ϵ les parties définies par la Figure 3.1 formant, avec Q_-^a et Q_-^b, une partition de $\mathbb{R}^2$. Soit enfin $u = \tan(\alpha)y_h$. On a*

– *si* $y^- \in Q_-$,

$$\frac{1}{2}\gamma_1([0;|y_v|]) + \frac{\alpha}{2\pi} + \gamma_1([0,\frac{y_v}{2}])\gamma_1\left(\left[0;\left|y_v/2\frac{\cos(\alpha)}{\sin(\alpha)}\right|\right]\right) \leq \gamma_2(Q_-^b - y^-)$$

$$\gamma_2(Q_-^b - y^-) \leq \frac{\alpha}{2\pi} + \gamma_1([0;|u|(1+\tan(\alpha))]), \tag{3.21}$$

– *si* $y^- \in Q_+$,

$$e^{-\frac{y_v^2}{2}}\frac{1}{2}\left(\frac{1}{2}\gamma_1([0;|u|]) + \frac{\alpha}{2\pi}\right) \leq \gamma_2(Q_-^b - y^-)$$

$$\gamma_2(Q_-^b - y^-) \leq e^{-\frac{\epsilon^2 y_v^2 \cos^2(\alpha)}{2(1+\epsilon)^2}}\left(\gamma_1([0;((1+\tan(\alpha))|u|]) + \frac{\alpha}{2\pi}\right), \tag{3.22}$$

– *si* $y^- \in Q_\epsilon$,

$$e^{-\frac{(1+\epsilon)^2|u|^2}{2}}\frac{1}{2}\left(\frac{1}{2}\gamma_1([0;|u|]) + \frac{\alpha}{2\pi}\right) \leq \gamma_2(Q_-^b - y^-)$$

$$\gamma_2(Q_-^b - y^-) \leq \left(\gamma_1([0;(1+\tan(\alpha))|u|]) + \frac{\alpha}{2\pi}\right). \tag{3.23}$$

– *On a concernant* $\gamma_2(Q_-^a - y^+)$ *:*

$$\gamma_2(Q_-^a - y^+) \leq \gamma_2(Q_-^b - y^-). \tag{3.24}$$

– *Enfin, si* $y_h = 0$*, on a*

$$e^{-\frac{y_v^2}{2}}\frac{\alpha}{2\pi} \leq \gamma_2(Q_-^a - y^+) = \gamma_2(Q_-^b - y^-). \tag{3.25}$$

La démonstration de ce lemme est reportée à la sous-section qui suit. Fixons pour le reste de la démonstration $\epsilon = 1$ dans le lemme précédent (les autres valeurs de ϵ servirons à la démonstration du Théorème 1.2). L'équation (3.24) du lemme implique que

$$\frac{1}{2}\gamma_2(Q_-^b - y^-) \leq \mathcal{R} \leq \gamma_2(Q_-^b - y^-).$$

Rappelons que (y_h, y_v) est défini à la suite de (3.19) comme les coordonnées de y^+ et que $u = \tan(\alpha)y_h$. Par ailleurs, un simple calcul permet d'obtenir

$$u = |d_0|\frac{\tan(\alpha)}{\|\Pi_{F_{10}^\perp}\hat{F}_{10}\|_{L_2(\gamma_C)}} \text{ et } y_v^2 = \frac{\|F_{10}\|_{L_2(\gamma_C)}^2}{4}.$$

Si $\frac{1}{2}|\langle G_p, \hat{G}_p\rangle_{\mathbb{R}^p}| < |d_0|$, on a dans le lemme précédent $y_- \in Q_-$ et :

$$\frac{1}{4}\gamma_1\left(\left[0;\frac{\tan(\alpha)\|F_{10}\|_{L_2(\gamma_C)}}{2}\right]\right) + \frac{\alpha}{4\pi} \leq \mathcal{R} \leq \frac{\alpha}{2\pi} + \gamma_1\left(\left[0;(1+\tan(\alpha))\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^\perp}\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right).$$

Le cas où $|d_0| < \frac{1}{4}|\langle G_p, \hat{G}_p\rangle_{\mathbb{R}^p}|$ (c'est-à-dire $2|u| < |y_v|$) correspond au cas où $y_- \in Q_+$, on a :

$$e^{-\frac{\|F_{10}\|_{L_2(\gamma_C)}^2}{8}}\frac{1}{4}\left(\frac{\alpha}{2\pi} + \frac{1}{2}\gamma_1\left(\left[0;\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^\perp}\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right)\right) \leq \mathcal{R}$$

et

$$\mathcal{R} \leq e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}\cos(\alpha)^2}{32}} \left(\frac{\alpha}{2\pi} + \gamma_1 \left(\left[0; (1+\tan(\alpha)) \frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^\perp} \hat{F}_{10}\|_{L_2(\gamma_C)}} \right] \right) \right).$$

Si $\frac{1}{4}|\langle G_p, \hat{G}_p\rangle_{\mathbb{R}^p}| < |d_0| < \frac{1}{2}|\langle G_p, \hat{G}_p\rangle_{\mathbb{R}^p}|$, (c'est-à-dire $2|u| > |y_v| > |u|$) on a dans le lemme précédent $y_- \in Q_\epsilon$ ($\epsilon = 1$), et puisque dans ce cas $|y_v| > |u| > |y_v|/2$, :

$$e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{2}} \frac{1}{4} \left(\frac{1}{2}\gamma_1 \left(\left[0; \frac{\|F_{10}\|_{L_2(\gamma_C)}}{4} \right] \right) + \frac{\alpha}{2\pi} \right) \leq \mathcal{R}$$

et

$$\mathcal{R} \leq \frac{\alpha}{2\pi} + \gamma_1 \left(\left[0; (1+\tan(\alpha)) \frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^\perp} \hat{F}_{10}\|_{L_2(\gamma_C)}} \right] \right)$$

Ceci achève la démonstration du Théorème 3.1. □

3.1.1 Démonstration du Théorème 1.2

Nous rappelons l'énoncé du Théorème 1.2 :

Théorème. *Supposons que* $0 < \alpha < \pi/2$ *(α défini par l'équation (1.8)), et que* $\cos(\alpha)\|F_{10}\|_{L_2(\gamma_C)} \to \infty$ *quand p tend vers l'infini. Alors :*

$$\mathcal{R} \to \begin{cases} 0 & \text{si } \liminf_{p\to\infty} \frac{2|d_0|}{|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|} < 1 \\ b \geq \frac{1}{8} & \text{si } \limsup_{n\to\infty} \frac{2|d_0|}{|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|} > 1 \end{cases} \quad \text{quand } p \to \infty.$$

Démonstration. Nous allons utiliser le lemme précédent en jouant sur la valeur de ϵ. Nous utilisons sans les rappeler les définitions données avant l'énoncé du lemme précédent. Plaçons nous dans le cas où $\frac{2|d_0|}{|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|}$ a une limite inférieur $a < 1$. Il existe alors $\epsilon > 0$ tel que y^+ et y^- (définis par (3.19)) appartiennent à Q_+ (à partir d'un certain rang) l'équation (3.22) implique que

$$\mathcal{R} \leq e^{-\frac{\epsilon^2\|F_{10}\|^2_{L_2}\cos^2(\alpha)}{2(1+\epsilon)^2}} \left(1 + \frac{|\alpha|}{2\pi} \right),$$

et $\mathcal{R}$ tend vers 0 lorsque $\|F_{10}\|^2_{L_2}\cos^2(\alpha)$ tend vers l'infini. Si $\frac{2|d_0|}{|\langle F_{10}, \hat{F}_{10}\rangle_{L_2(\gamma_C)}|}$ tend vers $a > 1$, alors y^+ ou y^- (défini par (3.19)) appartient à partir d'un certain rang à Q_-. Et puisque dans ce cas, d'après l'équation (3.21) du lemme précédent,

$$\mathcal{R} \geq \frac{1}{4} \left(\frac{1}{2}\gamma_1([0; \|F_{10}\|_{L_2}/2]) + \gamma_1 \left(\left[0; \frac{\|F_{10}\|_{L_2}\cos(\alpha)}{4\sin(\alpha)} \right] \right) \gamma_1([0; \|F_{10}\|_{L_2}/4]) + \frac{\alpha}{2\pi} \right), \quad (3.26)$$

on a le résultat voulu en faisant tendre $\|F_{10}\|_{L_2}$ vers l'infini. Il s'agit juste de remarquer que α dépend de $\|F_{10}\|_{L_2}$ et que les valeurs limites $\alpha = \pi/2$ et $\alpha = 0$ nécessitent d'utiliser des termes différents dans l'inégalité (3.26). Ceci achève la démonstration du Théorème 1.2. □

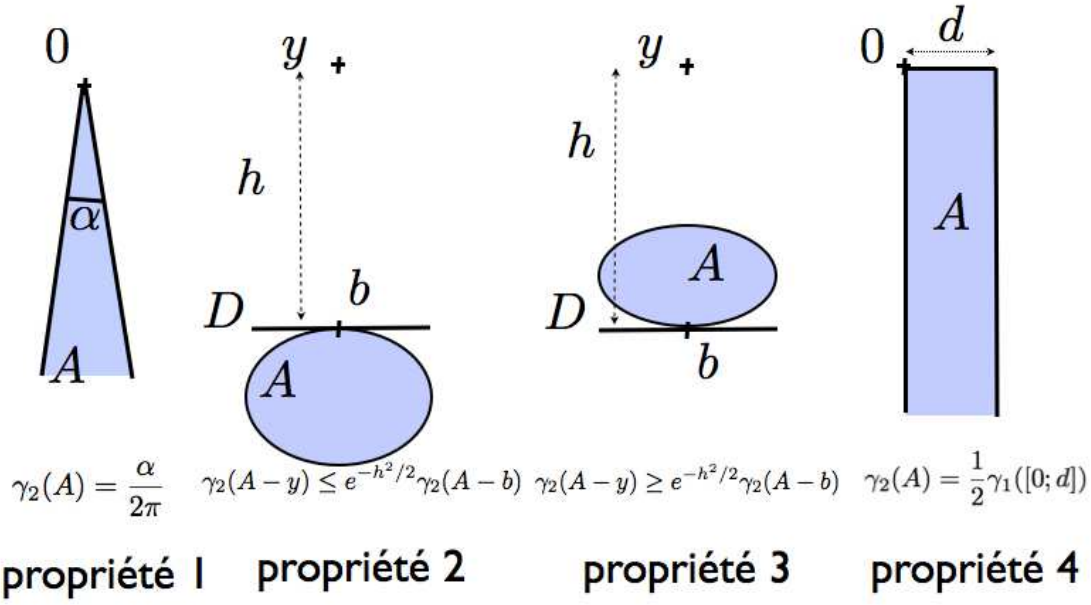


FIG. 3.2 – Les quatre propriétés utiles à la démonstration

3.1.2 Démonstration du Lemme 3.1

Cette démonstration est géométrique. Il s'agit simplement d'utiliser les quatre propriétés simples suivantes (illustrées par la Figure 3.2) :

- Propriété 1. Si A est une partie de $\mathbb{R}^2$ comprise entre deux demis-droites $(0,u)$ et $(0,v)$ telles que Angle$(u,v) = \alpha$, alors $\gamma_2(A) = \frac{\alpha}{2\pi}$. Ceci résulte directement de l'invariance par rotation de la mesure gaussienne. Une telle partie sera dite portion angulaire de taille α et de centre 0.
- Propriétés 2 et 3. Soient y un point de $\mathbb{R}^2$, D une droite de $\mathbb{R}^2$, b le projeté orthogonal de y sur D et h la distance de y à D. Si A est une partie de $\mathbb{R}^2$ incluse dans le demi-plan délimité par D et ne contenant pas y, alors $\gamma_2(A-y) \leq e^{-h^2/2}\gamma_2(A-b)$. C'est la propriété 2 . Si A est une partie de $\mathbb{R}^2$ incluse dans le demi-plan délimité par D et contenant y, alors $\gamma_2(A-y) \geq e^{-h^2/2}\gamma_2(A-b)$. C'est la propriété 3.
- Propriété 4. Si A est un rectangle de hauteur d et de largeur infinie (i.e A est du type $[0;d] \times [0;\infty[$ Figure 3.2) ayant 0 pour sommet alors $\gamma_2(A) = \frac{1}{2}\gamma_1([0;d])$. Un tel rectangle sera dit rectangle infini d'origine 0 et de hauteur d.

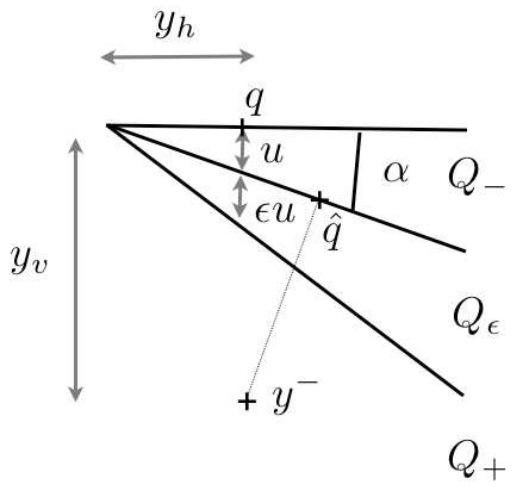


FIG. 3.3 – Figure de soutien pour la preuve

Nous noterons q et $\hat{q}$ les projetés orthogonaux de y sur D et $\hat{D}$. Les propriétés 2 et 3 sont bien connues, nous en rappelons cependant la démonstration. Il suffit de noter que

$$\gamma_2(A-y) = \int_{x\in A} \frac{1}{2\pi} e^{-\frac{\|x-y\|^2_{\mathbb{R}^2}}{2}} dx = e^{-\frac{h^2}{2}} \int_{x\in A} \frac{1}{2\pi} e^{-\frac{\|x-b\|^2_{\mathbb{R}^2}}{2}} e^{\langle x-b, y-b\rangle_{\mathbb{R}^2}} dx,$$

et que pour $x \in A$, $\langle x-b, y-b\rangle_{\mathbb{R}^2} \leq 0$ dans le cas de la propriété 2 et $\langle x-b, y-b\rangle_{\mathbb{R}^2} \geq 0$ dans le cas de la propriété 3.

Nous allons maintenant distinguer plusieurs cas, il ne s'agit plus que d'appliquer les 4 propriétés énoncées. Notons que l'inégalité concernant y^+ est évidente. Nous allons étudier $\gamma_2(Q^b_- - y^-)$ en distinguant plusieurs cas. On pourra s'aider des Figures 3.3 et 3.1.

cas $y^- \in Q^b_-$. Dans ce cas $|y_v| \leq |u|$. On peut inclure dans Q^b_- l'union disjointe d'un rectangle infini d'origine y^-, de hauteur $|y_v|$; d'une portion angulaire de taille α et de sommet y^- ; et d'un rectangle ayant y^- pour sommet, de hauteur $|y_v|/2$ et de largeur $|y_v/2\frac{\cos(\alpha)}{\sin(\alpha)}|$. En utilisant les propriétés 4 et 1, on a donc :

$$\frac{1}{2}\gamma_1([0;|y_v|]) + \frac{\alpha}{2\pi} + \gamma_1([0,\frac{y_v}{2}])\gamma_1\left(\left[0;\left|y_v/2\frac{\cos(\alpha)}{\sin(\alpha)}\right|\right]\right) \leq \gamma_2(Q^b_- - y^-). \tag{3.27}$$

D'autre part, Q^b_- peut être inclus dans l'union disjointe d'une section angulaire de centre y^-, de deux rectangles infinis de hauteur plus petite que $|u|\tan(\alpha)$ et de deux rectangles infinis de hauteur plus petite que $|u|$. Ainsi, les propriétés 1 et 4 impliquent :

$$\gamma_2(Q^b_- - y^-) \leq \frac{\alpha}{2\pi} + \gamma_1([0;|u|(1+\tan(\alpha))]). \tag{3.28}$$

cas $y^- \in Q_+$. Dans ce cas $|y_v| > (1+\epsilon)|u|$, y^- est à une distance $|y_v|$ de D et à une distance $(|y_v|-|u|)\cos(\alpha) \geq \frac{\epsilon}{1+\epsilon}|y_v|\cos(\alpha)$ de $\hat{D}$. Les propriétés 2 et 3 impliquent

$$e^{-\frac{y_v^2}{2}}\gamma_2(Q^b_- - q) \leq \gamma_2(Q^b_- - y^-) \leq e^{-\frac{\epsilon^2 y_v^2 \cos^2(\alpha)}{2(1+\epsilon)^2}}\gamma_2(Q^b_- - \hat{q}). \tag{3.29}$$

On peut inclure dans Q_-^b une section angulaire de taille α centrée en q ou un rectangle infini d'origine y et de hauteur $|u|$. Aussi, les propriétés 1 et 4 impliquent avec (3.29) et le fait que $\max(a,b) \geq \frac{a+b}{2}$ l'équation :

$$\frac{1}{2}\left(\frac{1}{2}\gamma_1([0;|u|]) + \frac{\alpha}{2\pi}\right) \leq \gamma_2(Q_-^b - q).$$

La partie Q_-^b peut être incluse dans l'union d'une section angulaire de taille α centrée en $\hat{q}$ et de deux rectangles infinis d'origine $\hat{q}$ et de hauteur $|u|(1+\tan(\alpha))$. Aussi, les propriétés 1 et 4 impliquent avec (3.29) et le fait que $\max(a,b) \geq \frac{a+b}{2}$ l'équation :

$$e^{-\frac{y_v^2}{2}}\frac{1}{2}\left(\frac{1}{2}\gamma_1([0;|u|]) + \frac{\alpha}{2\pi}\right) \leq \gamma_2(Q_-^b - y^-) \leq e^{-\frac{\epsilon^2 y_v^2 \cos^2(\alpha)}{2(1+\epsilon)^2}}\left(\gamma_1([0;|u|(1+\tan(\alpha))]) + \frac{\alpha}{2\pi}\right). \tag{3.30}$$

cas $y^- \in Q_\epsilon$. Dans ce cas $(1+\epsilon)|u| > |y_v| > |u|$, y^- est à une distance $|y_v| \leq (1+\epsilon)|u|$ de D et à une distance $(|y_v| - |u|)\cos(\alpha) \geq 0$ de $\hat{D}$. Les propriétés 2 et 3 impliquent

$$e^{-\frac{(1+\epsilon)^2|u|^2}{2}}\gamma_2(Q_-^b - q) \leq \gamma_2(Q_-^b - y^-) \leq \gamma_2(Q_-^b - \hat{q}). \tag{3.31}$$

on déduit l'inégalité suivante de la même manière que dans le sous-cas précédent :

$$e^{-\frac{(1+\epsilon)^2|u|^2}{2}}\frac{1}{2}\left(\frac{1}{2}\gamma_1\left([0;|u|]\right) + \frac{\alpha}{2\pi}\right) \leq \gamma_2(Q_-^b - y^-) \leq \left(\gamma_1([0;|u|(1+\tan(\alpha))]) + \frac{\alpha}{2\pi}\right). \tag{3.32}$$

Ceci termine la démonstration du lemme.

Remarque 3.1 (Sur les mesures log-concaves)**.** *Il est tout à fait naturel de s'interroger sur la validité des propriétés utilisées si la mesure γ n'est pas la mesure gaussienne. Pour ce qui est de la propriété 2, il est possible de considérer des mesures autres que gaussiennes. Supposons que μ soit une mesure de probabilité sur $\mathbb{R}^p$ avec une densité positive, $ae^{-\phi}$ par rapport à la mesure de Lebesgue, où ϕ est strictement convexe dans le sens ou il existe $c > 0$ tel que pour tout $x, y \in \mathbb{R}^p$*

$$\phi(x) + \phi(y) - 2\phi\left(\frac{x+y}{2}\right) \geq \frac{c}{2}\|x-y\|_{\mathbb{R}^p}^2, \tag{3.33}$$

$\phi(0) = 0 = \operatorname{Arginf}\phi$, a est une constante positive et ϕ est radiale : il existe une fonction ψ de $\mathbb{R}$ dans $\mathbb{R}$ telle que $\phi(x) = \psi(\|x\|)$. Soient y un point de $\mathbb{R}^p$, D un hyperplan de $\mathbb{R}^p$, b le projeté orthogonal de y sur D, h la distance de y à D et A une partie de $\mathbb{R}^p$ incluse dans le demi-espace délimité par D et ne contenant pas y.

Proposition 3.1. *Sous les conditions qui viennent d'être énoncées on a :*

$$\mu(A-y) \leq e^{-c\frac{h^2}{2}}\mu(A-b).$$

Démonstration. La démonstration se décompose en trois étapes.
Etape 1. Il suffit pour obtenir le résultat voulu, de démontrer que

$$\text{si } x, h \in \mathbb{R}^p \ \langle x, h\rangle_{\mathbb{R}^p} \geq 0 \text{ alors } \phi(x+h) - \phi(x) \geq \frac{c}{2}\|h\|^2_{\mathbb{R}^p}. \tag{3.34}$$

En effet, on a alors

$$e^{-\phi(x+h)} \leq e^{-\phi(x)} e^{-\frac{c}{2}\|h^2\|^2_{\mathbb{R}^p}},$$

Si $x \in A - b$ et $h = b - y$ on a : $\langle x, h\rangle_{\mathbb{R}^p} \geq 0$, et par intégration, l'équation précédente implique le résultat recherché :

$$\int_{A-b} e^{-\phi(x+y-b)} dx \leq \int_{A-b} e^{-\phi(x)} dx e^{-\frac{c}{2}\|h\|^2_{\mathbb{R}^p}}.$$

Par ailleurs, on peut supposer que ϕ est continue et différentiable.
Etape 2. Il suffit pour obtenir (3.34) de montrer que si $\langle x, h\rangle_{\mathbb{R}^p} \geq 0$, alors $\langle \nabla\phi(x), h\rangle_{\mathbb{R}^p} \geq 0$. En effet, puisque ϕ est fortement convexe et continue, on a pour tout $\theta \in [0, 1]$

$$\phi(x + \theta h) \leq \theta\phi(x) + (1-\theta)\phi(x+h) - \frac{c}{2}\theta(1-\theta)\|h\|^2_{\mathbb{R}^p},$$

et donc

$$\theta(\phi(x+h) - \phi(x)) \geq \frac{c}{2}\theta(1-\theta)\|h\|^2_{\mathbb{R}^p} + (\phi(x+\theta h) - \phi(x)).$$

Ainsi, en divisant par θ et en faisant tendre θ vers 0, on obtient (3.34) pourvu que $\langle \nabla\phi(x), h\rangle_{\mathbb{R}^p} \geq 0$.

Etape 3. Si $\langle x, h\rangle_{\mathbb{R}^p} \geq 0$, alors $\langle \nabla\phi(x), h\rangle_{\mathbb{R}^p} \geq 0$.
Puisque $\phi = \psi(\|x\|^2)$, $\nabla\phi(x) = 2\psi'(\|x\|^2)x$. Ainsi, $\nabla\phi(x)$ est colinéaire à x. Puisque ψ est minimal en 0, ce coefficient de colinéarité est positif. Ceci implique bien la propriété annoncée pour l'étape 3 et donc la proposition. □

3.1.3 Démonstration du Théorème 1.1

Théorème. *Soient $\hat{F}_{10}$ et $\hat{s}_{10}$ deux vecteurs de $\mathbb{R}^p$ et $\widehat{\mathcal{L}}^A_{10}(x)$ définie en substituant $\hat{F}_{10}$ et $\hat{s}_{10}$ à F_{10} et s_{10} dans (1.6). Soient P_1 et P_0 deux mesures gaussiennes sur $\mathcal{X} = \mathbb{R}^p$ de même covariance C et de moyennes respectivement μ_1 et μ_0. Alors, si $\hat{V}$ est une partie de $\mathbb{R}^p$ définie par (1.7), on a :*

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma)}} \quad \text{où } \mathcal{E} = \left(\frac{4\|F_{10}\|_{L_2(\gamma)}}{\sqrt{\pi}\|\hat{F}_{10}\|_{L_2(\gamma)}}|\langle \hat{F}_{10}, \hat{s}_{10} - s_{10}\rangle_{\mathbb{R}^p}| + \|F_{10} - \hat{F}_{10}\|_{L_2(\mathbb{R}^p, \gamma_C)}\right),$$

et $\mathcal{R}$ est l'erreur d'apprentissage donnée par la définition 5.1 Chapitre 5 Partie I. De plus, si $|\langle \hat{F}_{10}, \hat{s}_{10} - s_{10}\rangle_{\mathbb{R}^p}| \leq \frac{1}{4}|\langle \hat{F}_{10}, F_{10}\rangle_{L_2(\gamma_C)}|$ et $\langle \hat{F}_{10}, F_{10}\rangle_{L_2(\gamma_C)} \geq \frac{\sqrt{2}}{2}\|F_{10}\|_{L_2(\gamma_C)}\|\hat{F}_{10}\|_{L_2(\gamma_C)}$, alors

$$\mathcal{R}(\mathbb{1}_{\hat{V}}) \leq e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{32}} \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma)}}.$$

Démonstration. La dernière équation du théorème découle directement de l'équation (3.4) du Théorème 3.1. Pour la première, nous allons distinguer quatre cas.

1. Cas où $\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)} < 0$.
   Notons que puisque $\mathcal{R}_p$ est une probabilité, on a $\mathcal{R}_p \leq 1$. Par ailleurs,

   $$\mathcal{E} \geq \|F_{10} - \hat{F}_{10}\|_{L_2(\gamma_C)} \geq \|F_{10}\|_{L_2(\gamma_C)}.$$

   Ceci implique que $\mathcal{R}_p \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_C)}}$.

2. Cas où $\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)} > 0$ et $\|\hat{F}_{10}\|_{L_2(\gamma_C)} \leq \frac{1}{2}\|F_{10}\|_{L_2(\gamma_C)}$.
   Rappelons que $\mathcal{R}_p$ est borné supérieurement par $\frac{1}{2}$ lorsque $\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)} > 0$ (voir théorème 3.1, c'est le cas où α défini par (1.8) vérifie $-\pi/2 \leq \alpha \leq \pi/2$).
   Par ailleurs, puisque $\|\hat{F}_{10}\|_{L_2(\gamma_C)} \leq \frac{1}{2}\|F_{10}\|_{L_2(\gamma_C)}$, on a $\mathcal{E} \geq \frac{1}{2}\|F_{10}\|_{L_2(\gamma_C)}$, et ainsi, $\mathcal{R}_p \leq \frac{1}{2}$ implique que $\mathcal{R}_p \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_C)}}$.

3. Cas où $\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)} > 0$, $\|\hat{F}_{10}\|_{L_2(\gamma_C)} \geq \frac{1}{2}\|F_{10}\|_{L_2(\gamma_C)}$ et $\frac{\pi}{2} > \alpha > \frac{\pi}{4}$ (rappelons que α a été défini par 1.8).
   Puisque $\frac{\pi}{2} > \alpha > \frac{\pi}{4}$, on a $\cos(\alpha) \leq \frac{1}{2}$ et donc avec (1.8) :

   $$\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)} \leq \frac{\sqrt{2}}{2}\|\hat{F}_{10}\|_{L_2(\gamma_C)}\|F_{10}\|_{L_2(\gamma_C)}.$$

   Sous cette dernière contrainte, on a :

   $$\min_{\hat{F}_{10}} \|F_{10} - \hat{F_{10}}\|^2_{L_2(\gamma_C)} = \min_{\alpha} \left((1-\alpha)^2 + \alpha^2\right) \|F_{10}\|^2_{L_2(\gamma_C)} = \|F_{10}\|^2_{L_2(\gamma_C)},$$

   ce qui implique encore $\mathcal{R}_p \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_C)}}$.

4. Cas où $\langle \hat{F}_{10}, F_{10} \rangle_{L_2(\gamma_C)} > 0$, $\|\hat{F}_{10}\|_{L_2(\gamma_C)} \geq \frac{1}{2}\|F_{10}\|_{L_2(\gamma_C)}$ et $\alpha < \frac{\pi}{4}$.

   Puisque $\alpha \in [0, \frac{\pi}{4}]$, la concavité de la fonction sinus implique

   $$\frac{\alpha}{\pi} \leq \frac{\sin(\alpha)}{2\sqrt{2}}.$$

   D'autre part, puisque $\|\hat{F}_{10}\|_{L_2(\gamma_C)} \geq \frac{1}{2}\|F_{10}\|_{L_2(\gamma_C)}$, on a

   $$\sin(\alpha) = \frac{\|\Pi_{F_{10}^{\perp}} \hat{F}_{10}\|_{L_2(\gamma)}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}} \leq \frac{2\|F_{10} - \hat{F}_{10}\|_{L_2(\gamma)}}{\|F_{10}\|_{L_2(\gamma_C)}},$$

   (la première égalité est une formule de trigonométrie). Finalement cela donne

   $$\frac{\alpha}{\pi} \leq \frac{\|F_{10} - \hat{F}_{10}\|_{L_2(\gamma)}}{\sqrt{2}\|F_{10}\|_{L_2(\gamma_C)}}. \tag{3.35}$$

Rappelons que $d_0 = \langle \hat{F}_{10}, \hat{s}_{10} - s_{10} \rangle_{\mathbb{R}^p}$. L'égalité (1.8) définissant α, le fait que $\cos(\alpha) \geq \frac{\sqrt{2}}{2}$ impliquent maintenant :

$$\begin{aligned}\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^{\perp}}\hat{F}_{10}\|_{L_2(\gamma)}} &\leq \sqrt{2}|d_0|\frac{\sin(\alpha)}{\|\Pi_{F_{10}^{\perp}}\hat{F}_{10}\|_{L_2(\gamma)}} \text{ (puisque } \cos(\alpha) \geq \frac{\sqrt{2}}{2}) \\ &= \frac{\sqrt{2}|d_0|}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}} \text{ (d'après une formule de trigonométrie).}\end{aligned}$$

Ainsi, puisque $\gamma_1([0;u]) \leq \frac{u}{\sqrt{2\pi}}$, et que $\tan(\alpha) \leq 1$, on a :

$$\gamma_1\left(\left[0; (1+\tan(\alpha))\frac{|d_0|\tan(\alpha)}{\|\Pi_{F_{10}^{\perp}}\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right) \leq \gamma_1\left(\left[0; \frac{2\sqrt{2}|d_0|}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\right) \leq \frac{4|d_0|}{\sqrt{\pi}\|\hat{F}_{10}\|_{L_2(\gamma_C)}}. \tag{3.36}$$

Dans les cas 1, 2 et 3 du Théorème 3.1, puisque $\tan(\alpha) \leq 1$ ($\alpha \leq \frac{\pi}{4}$), les équations (3.35), (3.36) , (3.4),(3.7) impliquent :

$$\mathcal{R} \leq \frac{\mathcal{E}}{\|F_{10}\|_{L_2(\gamma_C)}}.$$

Ceci termine la démonstration du théorème.

□

3.1.4 Démonstration de la Proposition 1.1

Nous rappelons l'énoncé de cette proposition :

Proposition *Supposons que l'échantillon d'apprentissage est constitué de $n > p$ observations et que μ_1 et μ_0 sont connus. Soient $\hat{C}$ la covariance empirique et $\hat{C}^-$ l'inverse généralisée de Moore-Penrose de $\hat{C}$. Alors en prenant $\hat{F}_{10} = \hat{C}^- m_{10}$, on a*

$$\mathbb{E}_{P^{\otimes n}}[\mathcal{R}(\mathbb{1}_{\hat{V}})] \geq \frac{\arccos\left(\sqrt{\frac{n}{p}}\right)}{2\pi} e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}}.$$

Démonstration. La preuve est basée sur une idée donnée par Bickel et Levina [12] pour leur Théorème 1. Notons tout d'abord, comme le font Bickel et Levina, que si C est la matrice identité, il existe une base orthonormée dans $\mathbb{R}^p$ de vecteurs aléatoires de $\mathbb{R}^p$, $\xi_1, \ldots, \xi_p$, et un vecteur aléatoire $(\lambda_1, \ldots, \lambda_n)$ de $\mathbb{R}^n$ qui ont les propriétés suivantes.

1. Les λ_i sont indépendants entre eux, indépendants des $(\xi_i)_{i=1,\ldots,p}$, et $n\lambda_i$ suit une loi du χ^2 à $n-1$ degrés de liberté.
2. Pour tout i, ξ_i est tiré de manière indépendante et uniforme sur l'intersection de la sphère unité de $\mathbb{R}^p$ et de l'orthogonal à $\xi_1, \ldots, \xi_{i-1}$.
3. L'estimateur $\hat{C}$ empirique de C vérifie :

$$\hat{C} = \sum_{i=1}^{n} \lambda_i \xi_i \otimes \xi_i,$$

où si $x, y \in \mathbb{R}^p$, $x \otimes y$ est l'opérateur de $\mathbb{R}^p$ dans $\mathbb{R}^p$ qui à z associe $\langle x, z \rangle_{\mathbb{R}^p} y$.

Dans le cas général (C non nécessairement égal à I_p), on a γ_C-presque sûrement :

$$C^{-1/2}\hat{C}C^{-1/2} = \sum_{i=1}^{n} \lambda_i \xi_i \otimes \xi_i, \text{ et } C^{1/2}\hat{C}^{-}C^{1/2} = \sum_{i=1}^{n} \frac{1}{\lambda_i}\xi_i \otimes \xi_i.$$

Ainsi, en notant $\beta_i = \langle C^{-1/2}m_{10}, \xi_i\rangle^2_{\mathbb{R}^p}$, on a les équations suivantes :

$$\langle C^{-1}m_{10}, \hat{C}^{-}m_{10}\rangle_{L_2(\gamma_C)} = \langle C^{-1/2}m_{10}, C^{1/2}\hat{C}^{-}C^{1/2}C^{-1/2}m_{10}\rangle_{\mathbb{R}^p} = \sum_{i=1}^{n} \frac{\beta_i}{\lambda_i}, \tag{3.37}$$

$$\|\hat{F}_{10}\|^2_{L_2(\gamma_C)} = \sum_{i=1}^{n} \frac{\beta_i}{\lambda_i^2} \text{ et } \|F_{10}\|^2_{L_2(\gamma_C)} = \sum_{i=1}^{p} \beta_i. \tag{3.38}$$

Pour des raisons de symétrie (les ξ_i sont tirés de manière uniforme sur la sphère), on a pour tout sous-ensemble I_n de $\{1, \dots, p\}$ de taille n :

$$u_{I_n,p} = \mathbb{E}\left[\frac{\sum_{i\in I_n}\beta_i}{\sum_{i=1}^{p}\beta_i}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n}\beta_i}{\sum_{i=1}^{p}\beta_i}\right],$$

aussi, on obtient :

$$u_{I_n,p} = \frac{n}{p}. \tag{3.39}$$

Au vu des équations (3.37) et (3.38), l'espérance de l'angle α entre $\hat{F}_{10}$ et F_{10} dans $L_2(\gamma_C)$ (définit par 1.8) vaut

$$\begin{aligned}
\mathbb{E}[|\alpha|] &= \mathbb{E}\left[\arccos\left(\frac{\sum_{i=1}^{n}\frac{\beta_i}{\lambda_i}}{\sum_{i=1}^{p}\beta_i \sum_{i=1}^{n}\frac{\beta_i}{\lambda_i^2}}\right)\right] \text{ (définition de } \alpha) \\
&\geq \mathbb{E}\left[\arccos\left(\frac{\sum_{i=1}^{n}\beta_i}{\sum_{i=1}^{p}\beta_i}\right)\right] \\
&\quad \text{(inégalité de Cauchy-Schwartz et décroissance de la fonction arccos)} \\
&\geq \arccos\left(\mathbb{E}\left[\frac{\sum_{i=1}^{n}\beta_i}{\sum_{i=1}^{p}\beta_i}\right]\right) \text{ (inégalité de de Jensen et concavité de la fonction arccos sur } [0,1]) \\
&\geq \arccos\left(\sqrt{\frac{n}{p}}\right) \text{ (d'après 3.39).}
\end{aligned}$$

Au vu du Théorème 3.1 équation (3.8), ceci termine la preuve. □

3.1.5 Démonstration de la proposition 1.2

Proposition *Supposons que C est une matrice définie positive, que l'échantillon d'apprentissage est constitué de n observations, que $\hat{F}_{10} = C^{-1}(\bar{\mu}_1 - \bar{\mu}_0)$ et que $\hat{s}_{10} = s_{10}$. Alors, la règle de classification $\mathbb{1}_{\hat{V}}$ définie par (1.7) vérifie*

$$\mathbb{E}_{P^{\otimes n}}[\mathcal{R}(\mathbb{1}_{\hat{V}})] \geq \frac{\arccos\left(\frac{1}{\sqrt{p-2}}(\sqrt{n}\|F_{10}\|_{L_2(\gamma_C)} + 1)\right)}{2\pi} e^{-\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{8}}.$$

Démonstration. De la même manière que dans la proposition précédente, nous allons utiliser le Théorème 3.1 équation (3.8). Ainsi, il nous suffit, pour obtenir le résultat voulu de montrer que

$$\mathbb{E}\left[|\alpha|\right] \geq \arccos\left(\frac{1}{\sqrt{p-2}}(\sqrt{n}\|F_{10}\|_{L_2(\gamma_C)}+1)\right)$$

où α est l'angle entre $\hat{F}_{10}$ et F_{10} dans $L_2(\gamma_C)$ (défini par 1.8). Ainsi, puisque la fonction arccos est décroissante, concave sur $[0,1]$, il est suffisant d'obtenir l'inégalité

$$\mathbb{E}\left[\frac{|\langle F_{10},\hat{F}_{10}\rangle_{L_2(\gamma_C)}|}{\|F_{10}\|_{L_2(\gamma_C)}\|\hat{F}_{10}\|_{L_2(\gamma_C)}}\right] \leq \frac{1}{\sqrt{p-2}}(\sqrt{n}\|F_{10}\|_{L_2(\gamma_C)}+1). \tag{3.40}$$

Or,

$$\begin{aligned}\mathbb{E}\left[\frac{|\langle F_{10},\hat{F}_{10}\rangle_{L_2(\gamma_C)}|}{\|F_{10}\|_{L_2(\gamma_C)}\|\hat{F}_{10}\|_{L_2(\gamma_C)}}\right] &\leq \mathbb{E}\left[\frac{\|F_{10}\|_{L_2(\gamma_C)}}{\|\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]+\mathbb{E}\left[\frac{|\langle F_{10},\hat{F}_{10}-F_{10}\rangle_{L_2(\gamma_C)}|}{\|F_{10}\|_{L_2(\gamma_C)}\|\hat{F}_{10}\|_{L_2(\gamma_C)}}\right]\\ &\leq \mathbb{E}\left[\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{\|\hat{F}_{10}\|^2_{L_2(\gamma_C)}}\right]^{1/2}\left(1+\mathbb{E}\left[\frac{\langle F_{10},\hat{F}_{10}-F_{10}\rangle^2_{L_2(\gamma_C)}}{\|F_{10}\|^2_{L_2(\gamma_C)}}\right]^{1/2}\right),\end{aligned}$$

cette dernière inégalité résulte de l'inégalité de Cauchy-Scwartz. Rappelons que par définition de $\hat{F}_{10}$,

$$\hat{F}_{10} = F_{10}+\frac{C^{-1/2}}{\sqrt{n}}\xi,$$

où ξ est un vecteur gaussien centré réduit de $\mathbb{R}^p$. Ainsi, on obtient facilement, d'une part

$$\mathbb{E}\left[\frac{\langle F_{10},\hat{F}_{10}-F_{10}\rangle^2_{L_2(\gamma_C)}}{\|F_{10}\|^2_{L_2(\gamma_C)}}\right]^{1/2} = \frac{1}{\sqrt{n}},$$

et d'autre part :

$$\frac{\|F_{10}\|^2_{L_2(\gamma_C)}}{\|\hat{F}_{10}\|^2_{L_2(\gamma_C)}} = \frac{\|\sqrt{n}C^{1/2}F_{10}\|^2_{\mathbb{R}^p}}{\|\sqrt{n}C^{1/2}F_{10}+\xi\|^2_{\mathbb{R}^p}}$$

Le reste de la preuve repose sur le lemme suivant :

Lemme 3.2. *Soient $\sigma>0$, $\beta\in\mathbb{R}^p$, X une variable aléatoire gaussienne de $\mathbb{R}^p$ de moyenne β et de covariance I_p. Alors*

$$\mathbb{E}\left[\frac{1}{\|X\|^2_{\mathbb{R}^p}}\right] \leq \frac{1}{p-3}.$$

Pour démontrer ce lemme il suffit d'utiliser une base orthonormale $(e_i)_{i=1,\dots,p}$ de $\mathbb{R}^p$ obtenue en prenant $e_1=\beta/\|\beta\|_{\mathbb{R}^p}$. L'invariance par rotation de la mesure gaussienne implique que pour cette base, il existe $\xi_1,\dots,\xi_p$, p variables aléatoires indépendantes identiquement distribuées selon une loi normale centrée réduite, telles que

$$\|X\|^2_{\mathbb{R}^p} = (\xi_1+\|\beta\|_{\mathbb{R}^p})^2+\sum_{i=2}^p\xi_i^2 \geq \sum_{i=2}^p\xi_i^2.$$

Ainsi, nous avons

$$E\left[\frac{1}{\|X\|^2_{\mathbb{R}^p}}\right] \leq \mathbb{E}\left[\frac{1}{\sum_{i=2}^p \xi_i^2}\right].$$

Le résultat du lemme découle alors du fait que par un calcul direct (écrire l'intégrale avec la densité de $Z = \sum_{i=2}^p \xi_i^2$ un χ^2_{p-1} à $p-1$ degrés de liberté) on a :

$$\mathbb{E}\left[\frac{1}{\sum_{i=2}^p \xi_i^2}\right] = \frac{1}{p-2}.$$

□

3.2 Démonstration des résultats concernant la procédure QDA

3.2.1 Introduction

Exposé de la démarche. Dans cette section, nous allons démontrer les résultats concernant la procédure QDA. L'erreur d'apprentissage $\mathcal{R}$ (la probabilité de faire une erreur de classification avec la règle estimée alors que la règle optimale n'en fait pas) vérifie :

$$\mathcal{R} \leq \frac{1}{2}\left(P_1(X \in V_{\hat{\mathcal{L}}_{10}^Q} \triangle V_{\mathcal{L}_{10}^Q}) + P_0(X \in V_{\hat{\mathcal{L}}_{10}^Q} \triangle V_{\mathcal{L}_{10}^Q})\right) \tag{3.41}$$

(Si $f : \mathcal{X} \to \mathbb{R}$, V_f est défini par l'équation (3.1) au début de ce chapitre). En effet, l'événement $X \in V_{\hat{\mathcal{L}}_{10}^Q} \triangle V_{\mathcal{L}_{10}^Q}$ correspond au cas où les décisions (bonnes ou mauvaises) prises par la règle optimale et la règle estimée sont différentes.

Remarque 3.2. *Dans le cas de la procédure LDA, nous avions*

$$\mathcal{R} = \frac{1}{2}\left(\gamma_{C,s_{10}}\left(X \in \hat{V} \setminus V - \frac{m_{10}}{2}\right) + \gamma_{C,s_{10}}\left(X \in V \setminus \hat{V} + \frac{m_{10}}{2}\right)\right).$$

De cette équation, on déduit facilement (par des considérations de symétrie) que

$$2\mathcal{R} = \frac{1}{2}\left(\gamma_{C,s_{10}}\left(X \in \hat{V} \triangle V - \frac{m_{10}}{2}\right) + \gamma_{C,s_{10}}\left(X \in V \triangle \hat{V} + \frac{m_{10}}{2}\right)\right),$$

et donc que

$$2\mathcal{R} = \frac{1}{2}\left(P_1(X \in V_{\hat{\mathcal{L}}_{10}^A} \triangle V_{\mathcal{L}_{10}^A}) + P_0(X \in V_{\hat{\mathcal{L}}_{10}^A} \triangle V_{\mathcal{L}_{10}^A})\right). \tag{3.42}$$

On peut penser que ceci est encore valide dans le cas quadratique. Cela est bien moins évident (voir remarque 1.2 Chapitre 1 Partie I).

Dans la Sous-section 3.2.2 nous allons présenter une technique pour majorer des probabilités du type $P(V_f \triangle V_{f+\delta})$. Dans ce type de quantité nous appellerons perturbation la fonction mesurable δ (destinée à être petite) et fonction frontière optimale la fonction f mesurable de $\mathcal{X}$ dans $\mathbb{R}$. Dans le cas de la QDA, les résultats obtenus résultent tous du Théorème 3.2 énoncé ci-dessous, avec pour fonction frontière $f = \mathcal{L}_{10}^Q$ et pour perturbation $\delta = \hat{\mathcal{L}}_{10}^Q - \mathcal{L}_{10}^Q$.

Enoncé d'un théorème général concernant les perturbations quadratiques de règle quadratiques. Nous rappelons que $\mathcal{X}^*_{\gamma_{C,m}}$ est l'ensemble des applications affines mesurables sur $\mathcal{X}$ de carré intégrable et d'intégrale nulle par rapport à $\gamma_{C,m}$ et que $E_2(\gamma_{C,m})$ est l'espace des formes quadratiques mesurables d'intégrale nulle et de carré intégrable (voir définition Annexe B). En dimension infinie $H(\gamma_{C,m})$ est l'espace auto-reproduisant (voir Annexe B), en dimension finie, $\mathcal{X} = \mathbb{R}^p$, on a (si C est de rang plein) $H(\gamma_{C,m}) = \mathbb{R}^p$. Rappelons que à un opérateur symétrique A Hilbert-Schmidt sur $H(\gamma_{C,m})$ (en dimension finie une matrice symétrique), on associe la forme quadratique mesurable de $E_2(\gamma_{C,m})$ (voir Annexe B équation (B.6)). En dimension finie, si C est de rang plein :

$$\begin{aligned} q_A^{\gamma_{C,m}}(x) &= q_{C^{-1/2}AC^{-1/2}}(x-m) - \int_{\mathcal{X}} q_{C^{-1/2}AC^{-1/2}}(x-m)\gamma_{C,m}(dx) \\ &\quad (q_A(x) \text{ a été défini Chapitre 1 Partie I équation (1.11)}) \\ &= \langle AC^{-1/2}(x-m), C^{-1/2}(x-m)\rangle_{\mathbb{R}^p} - \sum_{i=1}^{p}\lambda_i, \end{aligned}$$

où $(\lambda_i)_{i=1,\ldots,p}$ est le vecteur des valeurs propres de A.

Theoreme 3.2. *Soient $\mathcal{X}$ un espace de Hilbert séparable, $\gamma_{C,m}$ la mesure gaussienne de covariance C et de moyenne m sur $\mathcal{X}$. Soient A et D, 2 opérateurs symétriques Hilbert-Schmidt sur $H(\gamma_{C,m})$, $F, d \in \mathcal{X}^*_{\gamma_{C,m}}$, et $c, d_0 \in \mathbb{R}$. Soient*

$$f(x) = c + F(x) + q_A^{\gamma_{C,m}}(x) \;\; et \;\; \delta(x) = d_0 + d(x) + q_D^{\gamma_{C,m}}(x)$$

les fonctions définissant V_f et $V_{f+\delta}$ (Si $g : \mathcal{X} \to \mathbb{R}$, V_g est défini par l'équation (3.1) au début de ce chapitre). Soit enfin $r \in \mathbb{R}$ tel que $r > 0$.

1. *Supposons que $r \leq \|f\|_{L_2(\gamma_{C,m})}$. Alors, pour tout $q \in]0,1[$, il existe $c_1(r,q) > 0$ (ne dépendant que de r et q) telle que*

$$\gamma_{C,m}(V_f \triangle V_{f+\delta}) \leq c_1(r,q)\|\delta\|^{q/3}_{L_2(\gamma_{C,m})}. \tag{3.43}$$

2. *Si $|\mathbb{E}_{L_2(\gamma_{C,m})}[f]| > r$, alors, pour tout $q \in]0,1[$, il existe $c_2(r,q) > 0$ (ne dépendant que de r et q) telle que*

$$\gamma_{C,m}(V_f \triangle V_{f+\delta}) \leq c_2(r,q)\|\delta\|^{2q/7}_{L_2(\gamma_{C,m})}. \tag{3.44}$$

Les deux sous-sections suivantes sont consacrées à la démonstration de ce résultat. La Sous-section 3.2.2 présente une méthodologie générale pour obtenir ce type de résultat et, dans la Section 3.2.4, nous appliquons cette méthodologie pour obtenir le Théorème 3.2.

3.2.2 Décomposition du domaine d'erreur

La différence symétrique entre deux parties A et B de $\mathcal{X}$ est notée $A\Delta B$. C'est l'ensemble des éléments $x \in \mathcal{X}$ qui sont dans un des deux ensembles mais pas dans les deux. Nous allons borner supérieurement la probabilité que X soit dans $V_f \Delta V_{f+\delta}$. Dans les cas que nous allons envisager, cet ensemble est composé essentiellement d'éléments pour lesquels soit δ prend de grandes valeurs, soit f est proche de zéro. Nous allons donc en même temps borner la mesure

1. des zones sur lesquelles la perturbation est grande (grâce à une inégalité de grande déviation)
2. et celle des zones sur lesquelles $|f|$ est petite (gâce à une inégalité du type $P(|f(X)| \leq \epsilon) \leq g(\epsilon)$).

C'est le rôle du Lemme 3.3 qui va suivre.

1. Hypothèse A_1. Il existe $c_0, c_1 > 0$, $h_\delta : \mathbb{R}^+ \to \mathbb{R}^+$ croissante telle que $h_\delta(0) = 0$, $\lim_{s\to\infty} h_\delta(s) = \infty$ et

$$\forall s > 0, \quad P\left(|\delta(X) - \mathbb{E}[\delta(X)]| \geq c_0 h_\delta(s)\right) \leq c_1 e^{-\frac{s^2}{2}}. \tag{3.45}$$

2. Hypothèse A_2. Il existe $\beta > 0$ et $c_2 > 0$ tels que

$$\forall \epsilon > 0, \quad P\left(|f(X)| \leq \epsilon\right) \leq c_2 \epsilon^\beta. \tag{3.46}$$

Remarque 3.3. *L'application h_δ de l'hypothèse A_1 va nous permettre de mesurer l'effet de la perturbation δ.*

Lemme 3.3. *Sous les hypothèses A_1 (3.45) et A_2 (3.46), pour tout $q \in]0;1[$ on a :*

$$\begin{aligned} P(X \in V_f \Delta V_{f+\delta}) \leq & c_1^{1-q} c_2 |\mathbb{E}_P[\delta(X)]|^{q\beta} \\ & + \sqrt{\frac{2\pi}{1-q}} \frac{c_2 c_1^{1-q}}{2} \mathbb{E}\left[\left(c_0 h_\delta\left(\frac{|\xi|}{\sqrt{1-q}} + 1\right) + |\mathbb{E}_P[\delta(X)]|\right)^{q\beta}\right], \end{aligned}$$

où ξ est une variable aléatoire gaussienne réelle centrée réduite.

Démonstration. Rappelons que $V_f = \{x : f(x) \geq 0\}$.

$$P(X \in V_f \Delta V_{f+\delta}) =$$

$$P\left(-(\delta(X) - \mathbb{E}[\delta(X)]) - \mathbb{E}[\delta(X)] \leq f(X) \leq 0 \text{ ou } 0 \leq f(X) \leq (\delta(X) - \mathbb{E}[\delta(X)]) + \mathbb{E}[\delta(X)]\right),$$

et donc

$$P(X \in V_f \Delta V_{f+\delta}) \leq P(U),$$

$$\text{où} \quad U = \left\{|f(X)| \leq |\delta(X) - \mathbb{E}[\delta(X)]| + |\mathbb{E}[\delta(X)]|\right\}.$$

Soient $B_j = \{c_0 h_\delta(j) \leq |\delta(X) - \mathbb{E}[\delta(X)]| < c_0 h_\delta(j+1)\}$ pour $j \in \mathbb{N}$. Cette famille d'évenement permet de recouvrir tous les événements possibles.

En écrivant que

$$P(U) = \sum_{j\geq 0} P(U \cap B_j),$$

puis en utilisant l'inégalité de Holder, ($p + q = 1$) on a :

$$P(U) \leq \sum_{j\geq 0} P(U \cap B_j)^q P(B_j)^p.$$

On obtient donc

$$P(X \in V_f \Delta V_{f+\delta}) \leq \sum_j P\left(|f(X)| \leq |\mathbb{E}[\delta(X)]| + c_0 h_\delta(j+1)\right)^q P\left(|\delta(X) - \mathbb{E}[\delta(X)]| \geq c_0 h_\delta(j)\right)^{1-q}$$
$$\leq c_2 c_1^{1-q} \sum_{j\geq 0} \left(|\mathbb{E}[\delta(X)]| + c_0 h_\delta(j+1)\right)^{q\beta} e^{-\frac{(1-q)j^2}{2}},$$
(d'après les hypothèses A1 et A2)

$$\leq c_2 c_1^{1-q} \left(|\mathbb{E}[\delta(X)]|^{q\beta_0} + \sqrt{\frac{2\pi}{1-q}} \int_0^\infty \left(h_\delta(x+1) + |\mathbb{E}[\delta(X)]|\right)^{q\beta} \sqrt{\frac{1-q}{2\pi}} e^{-\frac{(1-q)x^2}{2}} dx \right)$$

ce qui implique bien le résultat annoncé. □

Lemme 3.4. *Soient $\delta_1, \ldots, \delta_k$ k perturbations vérifiants l'hypothèse A_1 définie par l'équation (3.45) avec les fonctions d'erreur $h_{\delta_1}, \ldots, h_{\delta_k}$. Alors, si $h_\delta = \sum_{i=1}^k h_{\delta_i}$, il existe $c_0(k), c_1(k) > 0$ tels que*

$$\forall s > 0 \;\; P\left(|\delta - \mathbb{E}(\delta)| \geq c_0 h_\delta(s)\right) \leq c_1 e^{-\frac{s^2}{2}}. \tag{3.47}$$

Démonstration. Rappelons que pour tout i, $h_{\delta_i} \geq 0$. Fixons $s > 0$. La démonstration est basée sur le principe du nid de pigeon (pigeonhole principle). En effet, si $\sum_{i=1}^k |\delta_i - \mathbb{E}[\delta_i]| \geq k \sum_{i=1}^k c_{0i} h_{\delta_i}(s)$ alors il existe $i_0 \in \{1, \ldots, k\}$ tel que $|\delta_{i_0} - \mathbb{E}[\delta_{i_0}]| \geq \sum_{i=1}^k c_{0i} h_{\delta_i}(s)$. En posant $c_0 = k \max c_{0i}$, on a donc

$$P\left(\left|\sum_{i=1}^k \delta_i - \mathbb{E}[\delta_i]\right| \geq c_0 \sum_{i=1}^k h_{\delta_i}(s)\right) \leq P\left(\sum_{i=1}^k |\delta_i - \mathbb{E}[\delta_i]| \geq k \sum_{i=1}^k c_{0i} h_{\delta_i}(s)\right)$$
(d'après l'inégalité triangulaire et le fait que
$c_0 \sum_{i=1}^k h_{\delta_i}(s) \geq k \sum_{i=1}^k c_{0i} h_{\delta_i}(s)$)
$$\leq P\left(\exists i_0 \in \{1, \ldots, k\} \;:\; |\delta_{i_0} - \mathbb{E}[\delta_{i_0}]| \geq \sum_{i=1}^k c_{0i} h_{\delta_i}(s)\right)$$
(principe du nid de pigeon)
$$\leq \sum_{i=1}^k P\left(|\delta_i - \mathbb{E}[\delta_i]| \geq c_{0i} h_{\delta_i}(s)\right)$$
(sous additivité des probabilités)
$$\leq \sum_{i=1}^k c_{1i} e^{-\frac{s^2}{2}}$$
(h_{δ_i} vérifie l'hypothèse A_1),

ce qui achève la preuve. □

Les résultats permettant de vérifier l'hypothèse A2 sont présentés dans le section 3.3.1. Nous expliquons maintenant comment vérifier l'hypothèse A1.

3.2.3 Grandes déviations

L'hypothèse A_1 (3.45) est une inégalité de grande déviation sur la variable aléatoire réelle $\delta(X)$. Dans le cas où la perturbation est linéaire ou Lipschitz on peut utiliser le théorème suivant démontré par exemple dans le livre de Bogachev [15] (p174).

Theoreme 3.3. *Soient $\gamma = \gamma_C$ une mesure gaussienne de covariance C sur $\mathcal{X}$ un espace de Banach séparable , $H = H(\gamma)$ l'espace de Auto-reproduisant associé, $\delta : \mathcal{X} \to \mathbb{R}$ une application pour laquelle qu'il existe $N(\delta) > 0$ tel que :*

$$|\delta(x+h) - \delta(x)| \leq N(\delta)|h|_{H(\gamma)} \ \ \forall h \in H(\gamma) \ \ \gamma - ps \tag{3.48}$$

Alors

$$\forall s > 0 \ \ \gamma\left(x \in \mathcal{X} \ \ : |\delta(x) - \int \delta(x) d\gamma| > s\right) \leq 2e^{-\frac{s^2}{2N(\delta)^2}} \tag{3.49}$$

Notons que ce type d'inégalité n'est pas particulier à la mesure gaussienne (voir par exemple [50]). Dans le cas d'une perturbation quadratique, nous allons utiliser le résultat suivant de Massart et Laurent [49] (lemme 1 p1325).

Theoreme 3.4. *Si $D = Diag(d_1, \ldots, d_p)$ et $q_D(x) = \langle Dx, x\rangle_{\mathbb{R}^p}$, alors*

$$\gamma_{I_p,0}\left(x \in \mathbb{R}^p \ : \ q_D(x) - \int_{\mathbb{R}^p} q_D(x)\gamma_{I_p,0}(dx) \geq \frac{s}{2}\|q_D\|_{L_2(\gamma_{I_p,0})} + \sup_i |d_i| s^2\right) \leq e^{-\frac{s^2}{2}} \tag{3.50}$$

$$\gamma_{I_p,0}\left(x \in \mathbb{R}^p \ : \ q_D(x) - \int_{\mathbb{R}^p} q_D(x)\gamma_{I_p,0}(dx) \leq -\frac{s}{2}\|q_D\|_{L_2(\gamma_{I_p,0})}\right) \leq e^{-\frac{s^2}{2}} \tag{3.51}$$

Aussi, l'hypothèse A_1 est vérifiée pour $h_\delta(s) = \frac{s}{2}\|q_D\|_{L_2(\gamma_{I_p,0})} + s^2 \sup_i |d_i|) \leq \|q_D\|_{L_2(\gamma_{I_p,0})}(\frac{s}{2} + s^2)$.
L'utilisation que nous ferons des deux théorèmes précédent est contenu dans le corollaire suivant

Corollaire 3.1. *Soit $\mathcal{X}$ un espace de Banach séparable, γ une mesure gaussienne sur $\mathcal{X}$ et $\delta \in E_2(\gamma)$ (voir annexe B définition B.1 pour la définition de $E_2(\gamma)$). Alors δ vérifie l'hypothèse A1 avec $h_\delta(s) = \|\delta - \mathbb{E}_\gamma[\delta]\|_{L_2(\gamma)}(s + s^2)$*

Démonstration. Il suffit de montrer ce résultat pour $\mathcal{X} = \mathbb{R}^p$ et de procéder par un raisonnement d'approximation tel que celui effectué dans la preuve du théorème (1.4). Rappelons que dans $L_2(\gamma)$, on a $\mathcal{X}^*_{2,\gamma} = \{cte\} \oplus \mathcal{X}^*_\gamma \oplus E_2(\gamma)$ (voir annexe B définition B.1). Ainsi, il existe un unique triplet $\delta_0 = \mathbb{E}_\gamma[\delta] \in \{cte\}$, $\delta_1 \in \mathcal{X}^*_\gamma$ et $\delta_2 \in E_2(\gamma)$ tel que $\delta = \delta_0 + \delta_1 + \delta_2$. D'après le théorème précédent, l'hypothèse A_1 est vérifiée pour la perturbation δ_2, la mesure $P = \gamma$ et $h_{\delta_2}(s) = \|\delta_2\|_{L_2(\gamma)}(s + s^2)$. Puisque $\delta_1 \in \mathcal{X}^*_\gamma$, δ_1 est affine. Ainsi, d'après le Théorème 3.3, l'hypothèse $A1$ est vérifiée pour la perturbation δ_1 avec $h_{\delta_1}(s) = s\|\delta_1\|_{L_2(\gamma)}$. On peut alors conclure en utilisant le lemme 3.4 et le fait que

$$\|\delta_2\|_{L_2(\gamma)}(s + s^2) + s\|\delta_1\|_{L_2(\gamma)} \leq (\|\delta_1\|_{L_2(\gamma)} + \|\delta_2\|_{L_2(\gamma)})(s + s^2) \leq \sqrt{2}(s + s^2)\|\delta - \delta_0\|_{L_2(\gamma)}.$$

□

Nous avons maintenant tous les éléments pour démontrer le Théorème 3.2.

3.2.4 Démonstration du Théorème 3.2

Nous allons comme convenu appliquer le lemme 3.3. D'après le théorème 3.5 l'hypothèse $A2$ est vérifiée pour $\beta = 1/3$ dans le cas du point 1 du théorème et pour $\beta = 2/7$ dans le cas du point 2 du théorème. Dans les deux cas, la constante c_2 dépendant de r. Pour les deux points du théorème, d'après le corollaire précédent, l'hypothèse $A2$ est vérifiée avec la fonction $h_\delta(s) = (s + s^2)\|\delta - \delta_0\|_{L_2(\gamma)}$. Ainsi, si l'on applique le lemme 3.3, pour tout $q \in]0, 1[$, il existe une constante $C(r, q) > 0$ telle que

$$\gamma(V_f \Delta V_{f+\delta}) \leq C(r, q) \left(|\mathbb{E}_\gamma(\delta)| + \|\delta - \mathbb{E}[\delta]\|_{L_2(\gamma)} \right)^{q\beta},$$

et donc il existe une constante $C'(r, q) > 0$ telle que

$$\gamma(V_f \Delta V_{f+\delta}) \leq C'(r, q) \|\delta\|^{q\beta}_{L_2(\gamma)},$$

ce qui achève la démonstration.

3.3 Quelques lemmes techniques

3.3.1 Sur la densité d'une forme quadratique

Nous allons noter $l_2(\mathbb{N})$ l'ensemble des suites réelles de carrés sommables et $\mathcal{X}_2^*$ l'ensemble des variables aléatoires du type $c + \sum_{i\geq 1} \beta_i(\xi_i^2 - 1) + \alpha_i \xi_i$ où $c \in \mathbb{R}$, $\beta = (\beta_i)_i \in l_2(\mathbb{N})$, $\alpha = (\alpha_i)_i \in l^2(N)$ $(\xi_i)_{i\in\mathbb{N}}$ est une suite de variable aléatoires indépendantes identiquement distribuées de loi normale centrée réduite.

Soit $q \in \mathcal{X}_2^*$ donné par

$$q = c + \sum_{i\geq 0} \alpha_i \xi_i + \sum_i \beta_i(\xi_i^2 - 1).$$

Nous noterons

$$n_1(q) = \max_i |\alpha_i| \quad n_2(q) = \max_i |\beta_i|, \quad \sigma(q) = \left(\sum_{i\geq 0} 2\beta_i^2 + \alpha_i^2 \right)^{1/2}. \tag{3.52}$$

Theoreme 3.5. *1. Il existe une constante $C(c_0)$ telle que*

$$\sup \{ P(|q| \leq \epsilon) \ : \ q \in \mathcal{X}_2^* \ : \ |\mathbb{E}[q]| \geq c_0 \ \} \leq C(c_0)\epsilon^{2/7}.$$

2. Il existe une constante $C'(c_0)$ telle que

$$\sup \left\{ P(|q| \leq \epsilon) \ : \ q \in \mathcal{X}_2^* \ : \ E[q^2] \geq c_0 \ \right\} \leq C'(c_0)\epsilon^{1/3}.$$

3. Soit $q \in \mathcal{X}_2^*$, pour tout $\epsilon \geq 0$,

$$P(|q| \leq \epsilon) \leq \sqrt{\frac{1}{\pi} \frac{\epsilon}{n_2(q)}}.$$

(Notons que $\mathcal{X}_2^*$ est définit au début de cette section)

Remarque 3.4. *Ce résultat peut paraître surprenant, et nous ne sommes pas parvenu à montrer qu'il était optimal. Si* $n_2(q) = \max_i |\beta_i| > c_0$*, la borne du point* 3 *est optimale dans le sens où pour* $\beta = (1, 0, \dots)$*,* $c = 1$ *et* $\alpha = 0$ *on a* $P(|q| \leq \epsilon) = P(|\xi^2| \leq \epsilon) \sim C\epsilon^{1/2}$ *(pour une constante* C *que l'on peut calculer explicitement). Par ailleurs, lorsque* $\|\beta\|_{l^2} \to 0$ *le comportement de* $P(|q| \leq \epsilon)$ *tend à être celui de* $P(|\|\alpha\|_{l^2}\mathcal{N}(0,1) - c| \leq \epsilon) \sim C'(c_0)\epsilon$*. Ainsi, la conjecture selon laquelle les points* 1 *et* 2 *du théorème peuvent être améliorés (jusqu'à obtenir des exposants* $1/2$ *au lieu de* $2/7$ *et* $1/3$*) reste possible (même si nous ne pensons pas qu'elle soit vraie). Les cas difficiles à étudier (et le point* 3 *de la preuve qui va suivre en témoigne) sont les cas pour lesquels* $\|\beta\|_\infty \to 0$ *mais* $\|\beta\|_{l^2}$ *ne tend pas vers zéro.*

Démonstration. Nous allons procéder en plusieurs étapes

Etape 1. Nous allons montrer que si $|\mathbb{E}[q]| > \epsilon$ alors

$$P(|q| \leq \epsilon) \leq \frac{\sigma^2(q)}{(|\mathbb{E}[q]| - \epsilon)^2}. \tag{3.53}$$

Notons que $|q - \mathbb{E}[q]| \geq ||q| - |\mathbb{E}[q]||$ et si $|q| < \epsilon < |\mathbb{E}[q]|$ alors $||q| - |\mathbb{E}[q]|| = |\mathbb{E}[q]| - |q|$ et

$$|q| \geq |\mathbb{E}[q]| - |q - \mathbb{E}[q]|.$$

Ainsi

$$P(|q| \leq \epsilon) \leq P(|\mathbb{E}[q]| - |q - \mathbb{E}[q]| \leq \epsilon) = P(1 \leq \frac{|q - \mathbb{E}[q]|}{|\mathbb{E}[q]| - \epsilon})$$

ce qui implique (3.53) de par l'inégalité de Markov.

Etape 2. Notons tout de suite que l'on peut supposer sans restriction que pour tout $i \in \mathbb{N}$ $\alpha_i \geq 0$. C'est ce que nous ferons. Dans toute la suite de la preuve, $\alpha_{i_0} = \max_i \alpha_i$, $j_0 \in arg\max |\beta_j|$ et $\text{sign}(x)$ est la fonction qui donne le signe du réel x.

Nous allons montrer que

$$P(|q| \leq \epsilon) \leq \sqrt{\frac{1}{\pi}\frac{\epsilon}{n_2(q)}}, \tag{3.54}$$

Soit

$$Z = \sum_{i \neq j_0} \alpha_i \xi_i + \beta_i(\xi_i^2 - 1).$$

Pour obtenir cette inégalité, notons que pour tout $\alpha_{j_0} \geq 0$, $\beta_{j_0} \neq 0$

$$\begin{aligned} P\left(|Z + \alpha_{j_0}\xi + \beta_{j_0}(\xi^2 - 1)| \leq \epsilon\right) &= P\left(|\,\text{sign}(\beta_{j_0})Z + \alpha_{j_0}\xi + |\beta_{j_0}|(\xi^2 - 1)| \leq \epsilon\right) \\ &= P\left(|\frac{\text{sign}(\beta_{j_0})Z}{|\beta_{j_0}|} + (\xi + \frac{\alpha_{j_0}}{2|\beta_{j_0}|})^2 - 1 - \frac{\alpha_{j_0}^2}{4\beta_{j_0}^2})| \leq \frac{\epsilon}{|\beta_{j_0}|}\right) \\ &= P\left(\xi \in \left[f_{\alpha_{j_0},\beta_{j_0}}(-\epsilon) - \frac{\alpha_{j_0}}{2|\beta_{j_0}|}; f_{\alpha_{j_0},\beta_{j_0}}(\epsilon) - \frac{\alpha_{j_0}}{2|\beta_{j_0}|}\right]\right). \end{aligned}$$

où

$$f_{\alpha,\beta}(\epsilon) = \sqrt{(1 + \frac{\alpha^2}{4\beta^2} - \frac{\text{sign}(\beta)Z - \epsilon}{|\beta|})_+},$$

et $(x)_+ = x\mathbb{1}_{x\geq 0}$. L'inégalité (3.54) résulte alors du choix $\alpha = \alpha_{j_0}$ et $\beta = \beta_{j_0}$ et du fait que pour tout u, $\sqrt{(u+\frac{\epsilon}{|\beta_{j_0}|})_+} - \sqrt{(u-\frac{\epsilon}{|\beta_{j_0}|})_+} \leq \sqrt{\frac{2\epsilon}{n_2(q)}}$.

Etape 3 Nous allons montrer l'inégalité

$$P(|q| \leq \epsilon) \leq 208\frac{n_2(q)}{\sigma(q)} + \frac{2\epsilon}{\sigma(q)} e^{-\frac{(|\mathbb{E}[q]|-\epsilon)^2}{\sigma^2(q)}}. \tag{3.55}$$

Nous démontrons le Lemme suivant (qui est un théorème de la limite centrée) à la suite de la preuve.

Lemme 3.5. *Soient $X_i = \beta_i(\xi_i^2 - 1) + \alpha_i\xi_i$, ξ une variable aléatoire gaussienne centrée réduite et $\sigma(q)$ donné par (3.52) . On a :*

$$\sup_{\epsilon\geq 0}\left| P\left(|\mathbb{E}_\gamma[q] + \sum_{i\geq 0} X_i| \leq \epsilon\right) - P\left(|\xi + \frac{\mathbb{E}_\gamma[q]}{\sigma(q)}| \leq \frac{\epsilon}{\sigma(q)}\right)\right| \leq 104\frac{\max(|\beta_i|)}{\sigma(q)}.$$

Ainsi, puisque si $|\mathbb{E}[q]| > \epsilon$

$$P\left(|\xi + \frac{\mathbb{E}[q]}{\sigma(q)}| \leq \frac{\epsilon}{\sigma(q)}\right) \leq \frac{2\epsilon}{\sigma(q)} e^{-\frac{(|\mathbb{E}[q]|-\epsilon)^2}{\sigma^2(q)}},$$

on a bien l'inégalité (3.55).

Etape 4. Nous allons comme convenu distinguer plusieurs cas disjoints pour montrer les points 1 et deux 2 du théorème. Commençons par le point 1.

1. Dans le cas où $\sigma(q) < \epsilon^{1/7}$, c'est l'inégalité de l'étape 1 (3.53) qui permet de conclure.
2. Dans le cas où $n_2(q) \geq \epsilon^{3/7}$, c'est l'inégalité de l'étape 2 (3.54) qui permet de conclure.
3. Dans le cas où $n_2(q) < \epsilon^{3/7}$ et $\sigma(q) > \epsilon^{1/7}$, c'est l'inégalité de l'étape 3 (3.55) qui permet de conclure.

Terminons par le point 2.

1. Dans le cas où $n_2(q) \geq \epsilon^{1/3}$, c'est l'inégalité de l'étape 2 (3.54) qui permet de conclure.
2. Dans le cas où $n_2(q) < \epsilon^{1/3}$ c'est l'inégalité de l'étape 3 (3.55) qui permet de conclure.

□

Nous allons maintenant effectuer la preuve du Lemme 3.5.

Démonstration. Cette preuve est décomposée en deux étapes. Dans la première, nous effectuons le calcul de

$$\forall \alpha, \beta \in \mathbb{R}, \quad \phi_{\alpha,\beta}(t) = \mathbb{E}\left[e^{it(\xi\alpha + \beta(\xi^2-1))}\right], \tag{3.56}$$

et dans la deuxième étape nous en déduisons que pour tout $|t| < \frac{\sigma}{6\max_j|\beta_j|} = a$

$$|\prod_{j\geq 0} \phi_{\alpha_j,\beta_j}(t/\sigma) - e^{-t^2/2}| \leq \frac{4\max_j|\beta_j|}{\sigma}\frac{|t|^3}{2}e^{-t^2/6}, \tag{3.57}$$

ce qui implique bien le résultat voulu du fait de l'inégalité de Essen (voire par exemple [66] p358)

$$\begin{aligned}\sup_{u\in\mathbb{R}}\left|P\left(\frac{1}{\sigma}\sum_{j\geq 0}\alpha_j\xi_j+\beta_j(\xi_j^2-1)\geq u\right)-\Phi(u)\right| &\leq \int_{-a}^{a}\left|\frac{\prod_{i\geq 0}\phi_{\alpha,\beta}(t/\sigma)-e^{-t^2/2}}{t}\right|dt+\frac{24}{a\sqrt{2\pi}}\\ &\leq \frac{4\max_j|\beta_j|}{\sigma}\int_{\mathbb{R}}\frac{t^2}{2}e^{-\frac{t^2}{6}}dt+\frac{\max_j|\beta_j|72\sqrt{2}}{\sigma\sqrt{\pi}}\\ &= \frac{\max_j|\beta_j|}{\sigma}\left(72\sqrt{\frac{2}{\pi}}+32\right)\leq 104\frac{\max_j|\beta_j|}{\sigma},\end{aligned}$$

où Φ est la fonction de répartition d'une variable aléatoire réelle gaussienne centrée réduite.
Etape 1. Soient $\Omega_\beta=\{z\in\mathbb{C}\ \ 2\Im(z)\beta>-1\}$ et $\psi_{\alpha,\beta}(z)$ donné par

$$\forall\alpha,\beta\in\mathbb{R},\ \ z\in\omega_\beta\ \ \psi_{\alpha,\beta}(z)=\frac{e^{-\beta iz}}{(1-2\beta iz)^{1/2}}e^{-1/2\frac{\alpha^2z^2}{(1-2\beta iz)}}.$$

La fonction $\psi_{\alpha,\beta}$ est analytique sur Ω_β. La fonction $\phi_{\alpha,\beta}(t)$ définie par (3.56) peut être prolongée en une fonction analytique sur le dommaine Ω_β et puisque

$$\frac{x^2}{2}+y(\alpha x+\beta(x^2-1))=\frac{1}{2}(1+2\beta y)(x+\frac{\alpha y}{1+2\beta y})^2-\frac{\alpha^2y^2}{2(1+2\beta y)}$$

on observe que

$$\forall y>-\frac{1}{2\beta}\quad\psi_{\alpha,\beta}(iy)=\phi_{\alpha,\beta}(iy).$$

Ainsi, on en déduit que $\phi_{\alpha,\beta}(z)$ et $\psi_{\alpha,\beta}(z)$ coïncident sur Ω_β et donc en particulier sur $\mathbb{R}$ ce qui donne

$$\forall\alpha,\beta\in\mathbb{R},\ \ t\in\mathbb{R}\ \ \phi_{\alpha,\beta}(t)=\frac{e^{-\beta it}}{(1-2\beta it)^{1/2}}e^{-1/2\frac{\alpha^2t^2}{(1-2\beta it)}}.$$

Etape 2. Preuve de (3.57). L'équation précédente implique que

$$|\prod_{i\geq 0}\phi_{\alpha,\beta}(t/\sigma)-e^{-t^2/2}|=e^{-\frac{t^2}{2}}|e^z-1|\leq e^{-\frac{t^2}{2}}|z|e^z,$$

où

$$u=\frac{t}{\sigma}\ \text{ et }\ z=\frac{t^2}{2}+\sum_{j\geq 0}\left\{-1/2\frac{\alpha_j^2u^2}{(1-2\beta_jiu)}+\frac{1}{2}(-2\beta_jui-\log(1-2\beta_jui))\right\},$$

soit

$$z=\sum_{j\geq 0}\left\{\left(\frac{u^2\alpha_j^2}{2}-\frac{1}{2}\frac{\alpha_j^2u^2}{(1-2\beta_jiu)}\right)+\left(\frac{u^22\beta_j^2}{2}-\frac{1}{2}(2\beta_jui+\log(1-2\beta_jui))\right)\right\}. \tag{3.58}$$

D'une part, si $|t|<\frac{\sigma}{6\max_i|\beta_i|}$, alors pour tout $j\in\mathbb{N}$ $|2u\beta_j|<\frac{1}{3}$ et on a (cf développement de taylor équation (1) p352 dans [66])

$$|\log(1-2\beta_jui)+2\beta_jui-\frac{4\beta_j^2u^2}{2}|\leq\frac{8|u\beta_j|^3}{3}\left|\frac{1}{1-|2u\beta_j|}\right|\leq 4|u\beta_j|^2\max_j|\beta_j|.$$

D'autre part,

$$|\frac{u^2\alpha_j^2}{2} - \frac{1}{2}\frac{\alpha_j^2 u^2}{(1-2\beta_j iu)}| \leq \frac{1}{2}\alpha_j^2|u|^3\frac{2|\beta_j|}{1+4\beta_j^2u^2} \leq \alpha_j^2|u|^3\max_j|\beta_j|.$$

Aussi, si $|t| < \frac{\sigma}{6\max_i|\beta_i|}$, on a avec (3.58) :

$$|z| \leq 2\sigma^2|u|^3\max_j|\beta_j| = \frac{2\max_j|\beta_j|}{\sigma}|t|^3,$$

et

$$e^{-\left(\frac{t^2}{2}-|z|\right)} \leq e^{-\frac{t^2}{2}(1-\frac{2}{3})} = e^{-\frac{t^2}{6}}.$$

□

3.3.2 Calculs de variances et d'espérances

Lemme 3.6. *Supposons donné $\sum_{k=1}^K n_k$ variables aléatoires réelles gaussiens indépendantes dans $\mathbb{R}^p$:*

$$k = 1,\ldots,K \;\; i \in 1,\ldots,n_k \quad X_{ki} \rightsquigarrow \gamma_{0,\sigma_k} = \gamma^k,$$

et $(\alpha_{ij})_{(i,j)\in\{1\ldots,K\}^2}$ une famille de réels positifs tels que pour tout (i,j), $\alpha_{ij} = \alpha_{ji}$. La variable aléatoire

$$T = \sum_{i<j}\alpha_{ij}\sum_{u=1}^{n_i}\sum_{l=1}^{n_j}(X_{ui} - X_{lj})^2,$$

a pour espérance

$$\mathbb{E}[T] = \sum_{i=1}^K\sum_{j\neq i}\alpha_{ij}n_in_j\sigma_i^2,$$

et pour variance

$$Var[T] = 4\sum_{i=1}^K\sum_{j\neq i}\sum_{k\neq i}\alpha_{ik}\alpha_{ij}n_in_jn_k\sigma_i^4.$$

Démonstration. Puisque $\mathbb{E}[(X_{ui} - X_{lj})^2] = \sigma_i^2 + \sigma_j^2$, on déduit facilement l'espérance recherchée de :

$$\mathbb{E}_0[T] = \sum_{i<j}\alpha_{ij}n_in_j(\sigma_i^2+\sigma_j^2).$$

Nous allons maintenant nous tourner vers le calcul de la variance. Nous noterons $\delta_{ij} = 1_{i=j}$ $\delta_{ijk} = 1_{i=j=k}$. On a :

$$Var_0[T] =$$

$$\sum_{i_1<j_1}\sum_{h_1=1}^{n_{i_1}}\sum_{l_1=1}^{n_{j_1}}\sum_{i_2<j_2}\sum_{h_2=1}^{n_{i_2}}\sum_{l_2=1}^{n_{j_2}}\alpha_{i_1j_1}\alpha_{i_2j_2}E\left[(X_{h_1i_1} - X_{l_1j_1})^2(X_{h_2i_2} - Y_{l_2j_2})^2\right] - (\sigma_{i_1}^2+\sigma_{j_1}^2)(\sigma_{i_2}^2+\sigma_{j_2}^2). \tag{3.59}$$

Nous allons noter :

$$a_{i_1,j_1,i_2,j_2}^{h_1,l_1,h_2,l_2} = E\left[(X_{h_1i_1} - X_{l_1,j_1})^2(X_{h_2i_2} - X_{l_2,j_2})^2\right] - (\sigma_{i_1}^2+\sigma_{j_1}^2)(\sigma_{i_2}^2+\sigma_{j_2}^2).$$

Rappelons que si G_1, G_2, G_3, et G_4 sont quatre variables aléatoires gaussiennes, elles vérifient

$$\mathbb{E}[G_1G_2G_3G_4] = \mathbb{E}[G_1G_2]\mathbb{E}[G_3G_4] + \mathbb{E}[G_1G_3]\mathbb{E}[G_2G_4] + \mathbb{E}[G_1G_4]\mathbb{E}[G_3G_2].$$

On a donc

$$a^{h_1,l_1,h_2,l_2}_{i_1,j_1,i_2,j_2} = 2\mathbb{E}[(Y_{h_1i_1} - Y_{l_1,j_1})(Y_{h_2i_2} - Y_{l_2,j_2})]^2$$

$$= 2\left(\sigma^2_{i_1}\delta_{(i_1,h_1)(i_2,h_2)} + \sigma^2_{j_1}\delta_{(j_1,l_1)(j_2,l_2)} - \sigma^2_{i_1}\delta_{(i_1,h_1)(j_2,l_2)} - \sigma^2_{j_1}\delta_{(j_1,l_1)(i_2,h_2)}\right)^2$$

$$= 2(\sigma^4_{i_1}\delta_{(i_1,h_1)(i_2,h_2)}+\sigma^4_{j_1}\delta_{(j_1,l_1)(j_2,l_2)}+\sigma^4_{i_1}\delta_{(i_1,h_1)(j_2,l_2)}+\sigma^4_{j_1}\delta_{(j_1,l_1)(i_2,h_2)}+2\sigma^2_{i_1}\sigma^2_{j_1}\delta_{(i_1,h_1)(j_1,l_1)(i_2,j_2)(j_2,l_2)}$$
$$-\sigma^2_{i_1}\delta_{(i_1,h_1)(j_1,l_1)(i_2,h_2)} - \sigma^2_{i_1}\delta_{(i_1,h_1)(i_2,h_2)(j_2,l_2)}) - \sigma^2_{j_1}\delta_{(i_1,h_1)(j_1,l_1)(j_2,l_2)}) - \sigma^2_{j_1}\delta_{(j_1,l_1)(i_2,h_2)(j_2,l_2)}).$$

Ceci implique que :

$$a^{h_1,l_1,h_2,l_2}_{i_1,j_1,i_2,j_2}1_{i_1\neq j_1\ i_2\neq j_2} = 2\left(\sigma^4_{i_1}\delta_{(i_1,h_1)(i_2,h_2)} + \sigma^4_{j_1}\delta_{(j_1,l_1)(j_2,l_2)} + \sigma^4_{i_1}\delta_{(i_1,h_1)(j_2,l_2)} + \sigma^4_{j_1}\delta_{(j_1,l_1)(i_2,h_2)}\right)$$

On déduit de cette dernière expression et de (3.59) la variance :

$$Var_0[T] = 4\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\sum_{k=i+1}^{K}\alpha_{ij}\alpha_{ik}n_in_jn_k\sigma^4_i + 4\sum_{i=2}^{K}\sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\alpha_{ij}\alpha_{ik}n_in_jn_k\sigma^4_i,$$

dont le résultat final découle directement. □

Proposition 3.2. *Soient $n_1, \ldots, n_K$, K entiers, $X_1, \ldots, X_K$, K variables aléatoires indépendantes telles que $X_k \rightsquigarrow \chi^2_{n_k-1}$, et*

$$T = \sum_{i<j}\left(\frac{\sigma^2_i X_i}{n_i - 1} - \frac{\sigma^2_j X_j}{n_j - 1}\right)^2.$$

Alors, on a :

$$\mathbb{E}[T] = \sum_{i=1}^{K}\sum_{j\neq i}\frac{\sigma^4_j(n_i - 3)}{n_i - 1} - \sigma^2_i\sigma^2_j.$$

et

$$\mathrm{Var}(T)\left(1 + o\left(\max_{k=1,\ldots,K}\left(\frac{1}{n_k}\right)\right)\right) = 4\sum_{i<j}\left(\frac{\sigma^8_i}{n_i} + \frac{\sigma^8_i}{n_i} + \frac{(n_i + n_j)\sigma^4_i\sigma^4_j}{n_in_j}\right). \tag{3.60}$$

Démonstration. Commençons par noter que si X suit une loi du χ^2 à n degrés de liberté alors on a par un calcul simple :

$$\mathbb{E}[X^k] = \frac{2^k\Gamma\left(\frac{n+2k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \tag{3.61}$$

En particulier

$$\mathbb{E}[X] = n \text{ et } \mathbb{E}[X^2] = n(n+2). \tag{3.62}$$

D'autre part, T peut se réécrire de la manière suivante :

$$T = (K-1)\sum_{i=1}^{K}\frac{\sigma^4_i X^2_i}{(n_i - 1)^2} - \sum_{i\neq j}\frac{\sigma^2_i\sigma^2_j X_iX_j}{(n_i - 1)(n_j - 1)}. \tag{3.63}$$

Ainsi, on a :

$$\mathbb{E}[T] = \sum_{i=1}^{K} \sum_{j \neq i} \frac{\sigma_j^4 (n_i - 3)}{n_i - 1} - \sigma_i^2 \sigma_j^2.$$

Le calcul de la variance fait intervenir des termes trop complexes. Aussi nous en calculons une approximation en utilisant le fait que si $k, q < 4$, $i \neq j$ alors

$$\mathbb{E}[X_i^{k+q}] = \mathbb{E}[X_i^k X_j^q] \left(1 + o\left(\max\left(\frac{1}{n_i}, \frac{1}{n_j}\right)\right)\right). \tag{3.64}$$

Ainsi on peut supposer que T est une somme de termes indépendants pour calculer la variance de manière approchée. Posons pour cela :

$$V = \sum_{i<j} \mathrm{Var}\left(\left(\frac{\sigma_i^2 X_i}{n_i - 1} - \frac{\sigma_j^2 X_j}{n_j - 1}\right)^2\right).$$

On obtient facilement, grâce à (3.64) l'équation :

$$V = \mathrm{Var}(T) \left(1 + o\left(\max_{k=1,\ldots,K}\left(\frac{1}{n_k}\right)\right)\right). \tag{3.65}$$

Par ailleurs, on a :

$$\mathrm{Var}\left(\left(\frac{\sigma_i^2 X_i}{n_i - 1} - \frac{\sigma_j^2 X_j}{n_j - 1}\right)^2\right) \left(1 + o\left(\max\left(\frac{1}{n_i}, \frac{1}{n_j}\right)\right)\right)$$

$$= \sigma_i^8 \,\mathrm{Var}\left(\frac{X_i^2}{(n_i - 1)^2}\right) + \sigma_j^8 \,\mathrm{Var}\left(\frac{X_j^2}{(n_j - 1)^2}\right) + 4\sigma_i^4 \sigma_j^4 \,\mathrm{Var}\left(\frac{X_i X_j}{(n_i - 1)(n_j - 1)}\right).$$

D'une part

$$\mathrm{Var}(X_i X_j) = 2 n_i n_j (n_i + n_j) \left(1 + o\left(\max\left(\frac{1}{n_i}, \frac{1}{n_j}\right)\right)\right),$$

et d'autre part

$$\mathrm{Var}(X_i^2) = 4 n_i^3 (1 + o(\frac{1}{n_i})).$$

Finalement, on obtient bien :

$$\mathrm{Var}(T) \left(1 + o\left(\max\left(\frac{1}{n_i}, \frac{1}{n_j}\right)\right)\right) = 4 \sum_{i<j} \left(\frac{\sigma_i^8}{n_i} + \frac{\sigma_i^8}{n_i} + \frac{(n_i + n_j)\sigma_i^4 \sigma_j^4}{n_i n_j}\right).$$

□

Troisième partie

Segmentation d'images hyperspectrales

Dans cette partie, nous étudions le problème de segmentation d'images hyper-spectrales. Nous envisageons deux approches, l'une est supervisée et l'autre non supervisée. Le problème de segmentation d'image est un problème difficile dont les définitions peuvent être fluctuantes. Nous donnons au Chapitre 1 une définition du problème (celle donnée dans l'introduction de ce mémoire). Cette définition est en beaucoup de points semblable à la définition du problème de classification. La différence réside dans la similarité que l'on impose sur la probabilité d'apparition d'une classe donnée, en deux pixels voisins. Dans cette partie, l'observation en chaque pixel est une variable aléatoire à valeurs dans un espace de grande dimension.

Nous présentons au Chapitre 2 (de cette partie) une technique supervisée de segmentation. L'échantillon d'apprentissage nous permet de réduire la dimension du problème avec la procédure définie Section 3 Chapitre 2 Partie II. Ensuite, nous utilisons l'algorithme de Kolaczyk et al. [46] pour estimer les probabilités d'apparition des différentes classes. Nous terminons en utilisant une procédure de type plug-in. L'originalité de ce travail réside essentiellement dans la combinaison de ces trois étapes, et dans les résultats théoriques que nous avons obtenus.

Nous présentons au Chapitre 3 une technique non supervisée de segmentation. Cette technique repose sur l'utilisation de l'algorithme AWS (Adaptive Weigth Smoothing) de Polzehl et Spokoiny [59] pour construire une mesure de similarité entre pixels. Cet algorithme d'estimation mélange des procédures de test, une estimation à noyau adaptative et un algorithme de croissance de région. Nous adaptons les procédures de tests à la grande dimension des données grâce la technique de seuillage présentée Chapitre 3 Partie I. Ce seuillage constitue une réduction de dimension. La mesure de similarité entre les pixels est utilisée pour obtenir une segmentation. L'originalité de ce travail réside dans la combinaison d'une procédure de réduction de dimension et de l'algorithme AWS, et dans l'utilisation des poids de AWS pour produire une segmentation de l'image.

Chapitre 1

Problématique et modèle

> The mathematician, carried along on his flood of symbols, dealing apparently with purely formal truths, may still reach results of endless importance for our description of the physical universe.
>
> Pearson, Karl.

Dans ce chapitre, nous redonnons les définitions du problème de segmentation et du risque associé données en introduction de ce mémoire. Nous définissons les modèles de morceaux de frontières (ou modèles d'horizon), et une classe plus grande de fonctions constantes par morceaux. Nous définissons l'ensemble des partitions récursives dyadiques d'une image et donnons le lien entre ces partitions et les modèles de fonctions constantes par morceaux.

1.1 Le problème de segmentation

Le formalisme et la problèmatique de la segmentation d'images hyper-spectrales héritent de ceux de la classification en grande dimension. Rappelons qu'une image peut être modélisée par un ensemble structuré $\mathcal{T}_N$ de N pixels auxquels sont associées des observations $(x_i)_{i\in\mathcal{T}_N}$ à valeur dans un espace $\mathcal{X}$. On parle d'image hyper-spectrale lorsque $\mathcal{X}$ est un espace de grande dimension ou de dimension infinie. On parle d'image multidimensionnelle lorsque $\mathcal{X}$ est de dimension plus grande que 1 mais que cette dimension reste assez petite. Dans la suite, une image désignera de manière indifférente un de ces types d'images. La segmentation d'image consiste à prédire les labels $(y_i)_{i\in\mathcal{T}_N}$, associés aux observations $(x_i)_{i\in\mathcal{T}_N}$. Dans le cas le plus simple, celui de la segmentation binaire, un label prend ses valeurs dans $\{0,1\}$, et dans le cas de la segmentation à K classes, un label y_i prend ses valeurs dans $\{1,\dots,K\}$.

En segmentation à K classes, on construit une application h de $\mathcal{X}\times\mathcal{T}_N$ dans $\{1,\dots,K\}\times\mathcal{T}_N$ qui, à une observation $x\in\mathcal{X}$ en un pixel $i\in\mathcal{T}_N$, associe la prédiction faite. Cette application est

une fonction de décision que l'on appelle fonction de segmentation. Cette fonction de segmentation commet une erreur sur l'observation x au pixel i si $h(x,i) \neq y_i$.

Pour formaliser le problème de segmentation, de manière identique au problème de classification, il faut introduire un formalisme probabiliste. Ainsi, nous supposons que $(X_i, Y_i)_{i\in\mathcal{T}_N}$ est une famille de variables aléatoires indépendantes à valeurs dans $\mathcal{X}\times\{1,\dots,K\}$ modélisant les observations sur l'image et les classes associées. En un pixel $i \in \mathcal{T}_N$, la loi de X_i est notée P^i et nous notons P_k la loi de $(X_i|Y_i = k)$. Cette loi modélise les observations issues de la classe k, et, comme la notation l'indique, nous supposons que cette loi ne dépend pas de la position spatiale $i \in \mathcal{T}_N$. Nous souhaitons naturellement construire une fonction de segmentation performante, c'est-à-dire telle que l'espérance du nombre de pixels mal classés, à savoir :

$$\mathbb{E}\left[\sum_{i\in\mathcal{T}_N} 1_{h(X_i,i)\neq Y_i}\right], \tag{1.1}$$

soit la plus petite possible. La règle optimale, celle qui minimise cette espérance, est donnée par la règle de Bayes h^* qui s'exprime en fonction de $(P_i)_{i\in\mathcal{T}_N}$:

$$h^*(X_i,i) = \mathrm{Argmax}_k\, \pi_{ik}\frac{dP_k}{dP^i}(X_i). \tag{1.2}$$

Sans hypothèse supplémentaire, puisque les couples de variables $(X_i, Y_i)_{i\in\mathcal{T}_N}$ sont supposés indépendants, le problème de segmentation est exactement équivalent à N problèmes identiques de classification. L'intérêt d'un problème de segmentation est donné par la possibilité de rajouter une hypothèse modélisant la cohérence spatiale entre les différents pixels de l'image, et de mettre en oeuvre une procédure qui permet de tirer parti de cette cohérence. Dans ce but, nous supposons que l'ensemble $\mathcal{T}_N$ des pixels de l'image hyper-spectrale peut être découpé en M régions homogènes $\{A_1,\dots,A_M\}$. L'homogénéité de ces régions peut alors être modélisée en supposant que, dans une région donnée A_m, l'application qui à $i \in \mathcal{T}_N$ associe

$$\pi_{ik} = P(Y_i = k)$$

est constante. Notons que la connaissance des lois $(P_k)_k$ et des poids $(\pi_{ik})_{i\in\mathcal{T}_N,k\in\{1,\dots,K\}}$ permet de construire la règle de Bayes. Nous allons fabriquer une règle de segmentation de type plug-in, c'est-à-dire qui est construite de la même manière que la règle de Bayes, mais avec une estimation des poids $(\pi_{ik})_{i\in\mathcal{T}_N,k\in\{1,\dots,K\}}$ et des lois $(P_k)_k$. Pour réaliser notre estimation des poids et des densités des lois P_k, nous avons envisagé deux modélisations. Dans la première, nous n'imposons pas de restriction supplémentaire sur les poids $(\pi_{ik})_{i\in\mathcal{T}_N,k\in\{1,\dots,K\}}$ et supposons seulement que les régions $\{A_1,\dots,A_M\}$ sont la discrétisation de régions séparées par des frontières régulières. Ainsi, dans cette approche

$$P^i = \sum_{k=1}^{K}\pi_{ik}P_k \;\text{ et }\forall m\in\{1,\dots,M\},\;\; (i,j)\in A_m^2 \;\; P^i = P^j. \tag{1.3}$$

Dans la deuxième modélisation, on suppose que dans une zone A_m donnée, seulement une classe peut apparaître. En d'autres termes,

$$\forall m\in\{1,\dots,M\}\;\; \exists k\in\{1,\dots,K\} \;\; : \;\; \forall i\in A_m \;\; P(Y_i = k) = 1.$$

Dans les deux cas nous supposons que la loi P_k est donnée par une loi gaussienne sur un espace $\mathcal{X}$ de grande dimension. Dans le premier cas, les paramètres des lois $(P_k)_{k=1,\dots,K}$ sont estimés grâce à un échantillon d'apprentissage, et l'on parle alors de segmentation supervisée. Dans le deuxième cas, σ^2 est une constante d'échelle inconnue, la covariance des lois $(P_k)_{k=1,\dots,K}$ est $\sigma^2 I_d$ et les moyennes de ces lois sont inconnues. Dans cette deuxième approche, on ne possède pas d'échantillon d'apprentissage, et l'on ne connaît pas le nombre K de classes. On parle alors de segmentation non supervisée. Dans tous les cas, on construit une fonction de segmentation $\hat{h}$ que l'on compare avec la règle optimale h^* (définie par (1.2)). Pour cela, on peut définir, de la même manière que dans le cas de la classification, l'excès de risque par

$$\mathcal{S}(\hat{h}) = \mathbb{E}\left[\sum_{i\in\mathcal{T}_N} 1_{\hat{h}(X_i,i)\neq Y_i}\right] - \mathbb{E}\left[\sum_{i\in\mathcal{T}_N} 1_{h^*(X_i,i)\neq Y_i}\right]. \tag{1.4}$$

Notons (et ceci fait une différence avec la classification) que $\hat{h}(.,i)$ peut être obtenu à partir d'un échantillon d'apprentissage (composé de variables aléatoires indépendantes des observations X_i sur l'images), mais aussi à partir des observations X_i sur lesquels on veut effectuer la segmentation. Remarquons que dans le problème de segmentation tel que nous l'envisageons, nous tenons compte de la fréquence d'apparition des différentes classes alors que nous n'en tenions pas compte dans le problème de classification (cf Partie II). Rappelons que dans le problème de segmentation, cette fréquence modélise la présence dans l'image d'un type de spectre. La régularité imposée sur les fréquences d'apparitions des différentes classes est exactement ce dont nous cherchons à tirer parti.

1.2 Un modèle d'images

Nous allons utiliser, pour modéliser la régularité des différentes zones, un modèle de morceaux de frontières, aussi appelé modèle d'horizon, et un modèle un peu plus gros que celui-ci de fonctions constantes par morceaux. Nous les définissons dans un cadre continu. Nous terminons en définissant les partitions récursives dyadiques et en expliquant leur intérêt.
Dans toute la suite d est un entier supérieur ou égal à 2.

Remarque 1.1. *Notons que dans la suite une image sera une application de $[0,1]^d$ dans $\mathbb{R}$ avec $d \geq 2$. Attention, d n'a rien à voir avec la dimension spectrale. Dans nos applications, $d = 2$, mais nous voulons être en mesure de donner un algorithme et des résultats théoriques permettant d'analyser des images hyperspectrales « 3d » , c'est-à-dire des images sur lesquelles un spectre est observé pour chaque volume élémentaire de tout le cerveau (et non pas dans un plan de coupe du cerveau seulement). Dans ce cas, $d = 3$.*

1.2.1 Modèle d'horizon

Les modèles de « morceaux de frontières » ont été introduits par Korostelev et Tsybakov [47] et renommés par Donoho en modèle « d'horizon » (cf [25]). Ils visent à modéliser des images. L'image y est définie comme deux zones séparées par le graphe d'une fonction sur laquelle des hypothèses de régularité sont faites. Rappelons que $g : [0,1]^{d-1} \to [0,1]$ est Lipschitzienne de

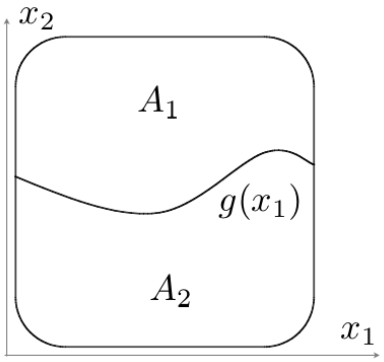


FIG. 1.1 – Modèle d'horizon pour $d = 2$

constante C ($g \in \mathcal{L}(C)$) lorsque

$$\forall x, x' \in [0,1]^{d-1}, \quad |g(x) - g(x')| \leq C\|x - x'\|.$$

Les modèles d'horizon sont définis par $\mathcal{H}_d(C, M)$ l'ensemble des fonctions $f : [0,1]^d \to \mathbb{R}$ pour lesquelles $\|f\|_\infty \leq M$ et il existe une fonction $g : [0;1]^{d-1} \to [0;1]$ Lipschitzienne de constante C, appelée fonction frontière, telle que :

$$A_1 = \{x = (x_1; x_2) \in [0,1]^{d-1} \times [0,1] \; tq \; x_2 \geq g(x_1)\} \;\text{ et }\; f(x) = a_1 1_{A_1}(x) + a_2 1_{A_2}(x), \quad (1.5)$$

où a_1 et a_2 sont deux éléments distincts de $\mathbb{R}$. Supposer qu'une image appartient à un modèle d'horizon peut paraître au premier abord, un peu restrictif. En fait, localement, beaucoup d'images sont incluses dans les modèles d'horizon. C'est ce qui justifiait le nom donné à ces modèles par Tsybakov et Korostelev : « morceaux de frontière » . Notons qu'il est possible de supposer d'autres types de régularités sur g (par exemple $g \in C^k$ ou $g \in \mathcal{B}_{s',p,q}$...).

Les propriétés des modèles d'horizon se traduisent par une relation simple. Le fait que la frontière g soit lipschitzienne implique une borne sur sa variation dans un pixel. En effet, si chaque pixel est un hypercube $I_j \times [\frac{i}{n}; \frac{i+1}{n}]$ de coté $1/n$ et que $g \in \mathcal{L}(C)$, on a la majoration suivante :

$$\sup_{x,y \in I_j} |g(x) - g(y)| \leq C\sqrt{d-1}\frac{1}{n}. \quad (1.6)$$

1.2.2 Partition récursive dyadique.

Les partitions récursives dyadiques (RDP) sont un outil algorithmique et théorique. Elle nous serviront au chapitre suivant. Nous en rappelons ici la définition (donnée à partir d'un quadtree) et rappelons un résultat qui fait le lien entre modèle d'horizon et RDP.

Rappelons qu'un 2^d-tree est un arbre dans lequel chaque noeud donne lieu à 2^d branches ou 2^d feuilles. Une partition dyadique récursive $\mathcal{P}(\mathcal{T})$ de $[0,1]^d$ associée à un 2^d-tree $\mathcal{T}$ est une partition construite par l'appel $RDP_d([0,1]^d, \mathcal{T})$ de la procédure récursive $RDP_d(I, \mathcal{T})$. Cette dernière procédure a pour argument un hypercube $I \subset [0,1]^d$ et un 2^d-tree $\mathcal{T}$. Elle est définie comme suit.

– **Initialisation**. $\mathcal{P} = I$, et $x = s_0$ la racine de l'arbre $\mathcal{T}$.

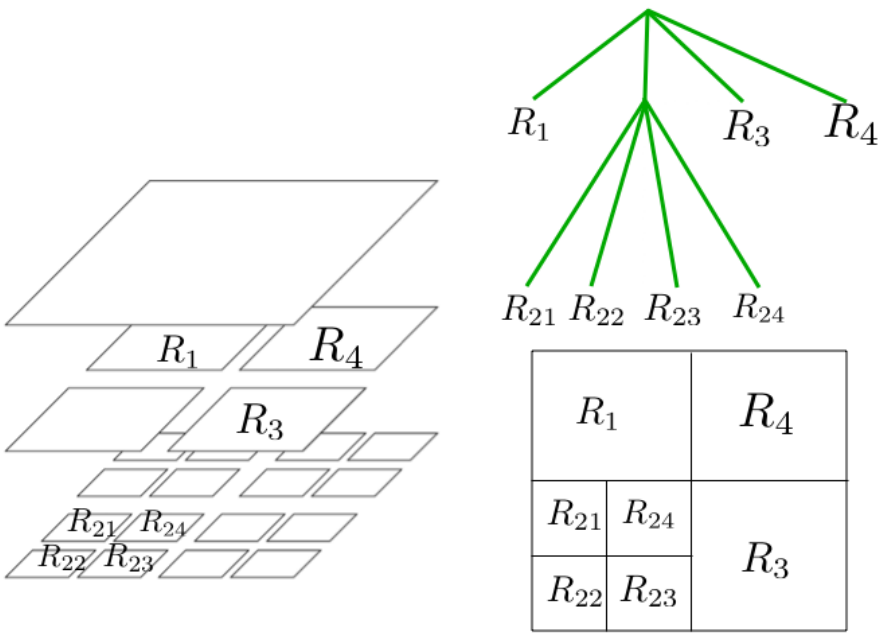


FIG. 1.2 – Construction d'une RDP de $[0,1]^2$ à partir d'un quadtree (4-tree) et vision multiéchelle

– **Appel récursif**.
Si x n'est pas une feuille de $\mathcal{T}$, alors
$\rightarrow$ Soient $(\mathcal{T}_i)_{i=1,\dots,2^d}$ les 2^d sous-arbres de $\mathcal{T}$ issus de x
(ceux dont les racines respectives sont les quatre fils de x).
Soit $(I_i)_{i=1,\dots,2^d}$ la partition en 2^d hypercubes égaux de I.
Faire $\mathcal{P} = \cup_{i=1}^{2^d} RDP_d(I_i, \mathcal{T}_i)$.
Si x est une feuille retourner $RDP_d(I, \mathcal{T}) = \{I\}$.

La figure 1.2 illustre cette construction dans le cas $d = 2$. L'ensemble des RDP de $[0,1]^d$ est défini comme étant l'ensemble des partitions $\mathcal{P}$ pour lesquelles il existe un 2^d-tree $\mathcal{T}$ tel que $\mathcal{P} = RDP_d([0,1]^d, \mathcal{T})$

1.2.3 Modèle de fonctions constantes par morceau

Définition 1.1. *Soient $f : [0,1]^d \rightarrow \mathbb{R}$ une fonction constante par morceaux $B(f)$ l'ensemble des points de discontinuité de f. Soit $N(f,r)$ le nombre de minimal d'hypercubes d'une RDP de longueur r permettant de recouvrir $B(f)$. L'ensemble des fonctions constantes par morceaux de paramètre (β, M) est défini par*

$$CM_d(\beta, M) = \left\{ f : [0,1]^d \rightarrow \mathbb{R} \ : \ \|f\|_\infty \leq M \quad et\ N(f,r) \leq \beta r^{-(d-1)} \right\}.$$

D'après (1.6), $\mathcal{H}_d(\sqrt{d-1}\beta, M) \subset CM_d(\beta, M)$. Ainsi, dans la suite nous nous référerons à $CM_d(\beta, M)$ tout en gardant à l'esprit que ce qui est valable pour $CM_d(\beta, M)$ l'est aussi pour $\mathcal{H}_d(\sqrt{d-1}\beta, M)$.

La proposition qui suit a été le point de départ de certains travaux de Donoho (cf [25] et [24]). Elle fait le lien entre les RDP et les modèles d'horizon. Les RDP correspondent à une famille de partitions permettant d'obtenir une bonne approximation d'images dans les modèles d'horizon avec assez peu d'éléments (on peut parler d'un bon rapport richesse/complexité).

Proposition 1.1. *Soient $\beta, M > 0$, et $H \in CM_d(\beta, M)$.*

1. *Il existe une partition dyadique récursive $\mathcal{P}$, de cardinal borné par m^{d-1}, telle que la meilleure approximation en norme L_1 $\bar{H}$ de H par des fonctions constantes sur les éléments de $\mathcal{P}$ vérifie*

$$\|H - \bar{H}\|_{L_1} \leq C(d, M, \beta)\frac{1}{m}, \;\; C(d, M, \beta) = 2^{\frac{2d-1}{d-1}} M \beta^{\frac{d}{d-1}} \tag{1.7}$$

2. *Soit PI_N la partition régulière de $[0,1]^d$ en N hypercubes identiques (i.e N pixels). Il existe une partition récursive dyadique $P(m)$, de taille m^{d-1}, telle que le nombre d'éléments de PI_N inclus dans un élément de $P(m)$ ayant une intersection avec la frontière de H ne soit pas plus grand que $c\frac{N}{m}$ (c étant une constante positive indépendante de N et m).*

Démonstration. Nous confondons dans cette démonstration une partition dyadique récursive et le 2^d-tree qui lui est associé. Puisque $N(f, r) \leq \beta r^{-(d-1)}$, il est possible de recouvrir $B(H)$ (l'ensemble des points de discontinuité de H) par un ensemble $B_j(H)$ d'hypercubes de longueur 2^{-j} et tel que $|B_j| \leq \beta 2^{j(d-1)}$. Il est clair que tous les éléments de B_j sont nécessairement des enfants d'éléments de B_{j-1}. Aussi, il est possible de choisir $2^{(j-1)(d-1)}\beta$ parents aux éléments de B_j lesquels ont en tout $2^d 2^{(j-1)(d-1)}\beta$ enfants.

On construit maintenant $\mathcal{P}$ (plus précisement le 2^d-tree associé) de la manière suivante. On raisonne de manière récursive avec pour donnée initiale un arbre composé d'une feuille. On remplace de manière récursive une feuille par 2^d branches lorsque l'hypercube correspondant à cette feuille a une intersection non vide avec $B(H)$. Finalement si l'on effectue cette procédure jusqu'à l'échelle J, le cardinal de $\mathcal{P}$ est borné par

$$\sum_{j \leq J} 2^{d+(j-1)(d-1)}\beta \leq 2^d \beta 2^{J(d-1)} = m_+^{d-1}. \tag{1.8}$$

A cette échelle J, notons $\bar{H}_J$ la meilleure approximation (pour la norme $\|.\|_{L_1}$) de H par une fonction constante par morceaux sur les éléments de $\mathcal{P}$. Il est clair qu'elle ne diffère de H que sur les éléments de la partition $\mathcal{P}$ qui intersectent $B(H)$. Ceux-ci ne sont pas plus que $\beta 2^{J(d-1)}$ et chacun d'entre eux a une mesure de Lebesgue de 2^{-dJ}. On a donc

$$\|\bar{H}_J - H\|_1 \leq 2M\beta 2^{J(d-1)} 2^{-dJ} = 2M\beta 2^{-J} = \frac{C(d, M, \beta)}{m_+},$$

ce qui termine la preuve. □

Chapitre 2

Méthode multi-échelle

Toute vision se change en contemplation, toute contemplation en réflexion, toute réflexion en association, de sorte que l'on peut dire que chaque fois que nous jetons un regard attentif sur le monde, nous faisons déjà de la théorie.

Goethe.

Dans ce chapitre, nous donnons un algorithme de segmentation supervisée. Nous l'appliquons à des données médicales. Nous présentons et démontrons les résultats théoriques que nous avons obtenus.

2.1 Fonction de segmentation et résultat principal

L'objectif de ce chapitre est d'exposer une méthode de segmentation d'images hyperspectrales. Cette méthode est basée sur la modélisation du chapitre précédent. On a à disposition deux jeux de données :

1. Les observations sur l'image que l'on veut segmenter : $(X_i)_{i\in\mathcal{T}_N}$. Les X_i sont des réalisations indépendantes et pour un pixel $i \in \mathcal{T}_N$, la loi de X_i est

$$P^i = \sum_{k=1}^{K} \pi_{ik} P_k.$$

2. Un échantillon d'apprentissage constitué pour $k \in \{1, \ldots, K\}$ de $(Z_i^k)_{i\in 1,\ldots,n_k}$, n_k réalisations indépendantes d'une variable aléatoire de loi P_k.

Hypothèses. Nous allons faire les hypothèses suivantes

Hypothèse **A0**-d (Image d-dimensionnelle régulière, $d \geq 2$). Soit PI_N la partition régulière de $[0,1]^d$ en N hypercubes identiques (i.e N pixels). Pour tout $k \in \{1, \ldots, K\}$, il existe $\beta > 0$, $M < \infty$ et $f_k \in CM_d(\beta, M)$ (voir chapitre précédent définition (1.1)) tels que

$$\forall i \in \mathcal{T}_N, \quad \pi_{ik} = f_k(t_i), \tag{2.1}$$

où t_i est le centre de l'hypercube i de PI_d.

Remarque 2.1. *Cette hypothèse est une hypothèse sur la structure topologique de l'ensemble des pixels $\mathcal{T}_N$. Cette structure est d'autant plus complexe que d est grand.*

Hypothèse **A1**. Il existe une constante positive B telle que

$$\sup_{x \in \mathcal{X}, k_1, k_2 \in \{1,\ldots,K\}^2} \frac{dP_{k_1}}{dP_{k_2}}(x) \leq B.$$

Remarque 2.2. *Cette dernière hypothèse est nécessaire pour obtenir des résultats théoriques. Elle est omniprésente dans les techniques d'estimation dans les modèles de mélanges (voir par exemple la thèse de Li [52]). Notons seulement que si P_1 et P_2 sont deux mesures gaussiennes équivalentes alors, la dérivée de Radon-Nykodym entre ces deux mesures est presque sûrement finie (cf [15]) mais pas bornée. Dans l'application de cet algorithme, nous utiliserons tout de même des distributions gaussiennes. Pour faire en sorte que l'Hypothèse **A1** soit vérifiée, on peut tronquer ces distributions à un même support lorsque celles-ci sont proches (au sens de la distance L_1) et les tronquer sur des supports disjoints lorsqu'elles sont "éloignées". Dans ce dernier cas l'Hypothèse **A1** n'est plus vérifiée, mais les distributions tronquées sont orthogonales et on peut penser que la détection est parfaite.*

Algorithme et résultat théorique. L'algorithme se décompose en trois phases

1. Utiliser la méthode donnée à la Section 3 du Chapitre 2 de la Partie II de ce mémoire pour construire un estimateur $\hat{P}_k$ de la loi P_k (alors supposée gaussienne) à partir des observations $(Z_i^k)_{i,k}$.
2. Supposer que les lois P_k sont connues égales à $\hat{P}_k$. Utiliser l'algorithme de Kolaczyk et al. [46] rappelé à la sous-section 2.2 pour construire un estimateur $(\hat{\pi}_{ik})_{ik}$ du vecteur de poids $(\pi_{ik})_{ik}$ à partir des observations sur l'image $(X_i)_{i \in \mathcal{T}_N}$.
3. Utiliser la fonction de segmentation (de type plug-in) définie par

$$\hat{h}(X_i, i) = \mathrm{Argmax}_k \, \hat{\pi}_{ik} \frac{dP_k}{d\hat{P}^i}(X_i) \quad \text{où} \quad \hat{P}^i = \sum_{k=1}^{K} \hat{\pi}_{ik} P_k. \tag{2.2}$$

Nous avons obtenu pour cette fonction de segmentation le théorème suivant.

Theoreme 2.1. *Soit $d \geq 2$. Supposons que $K = 2$ et que pour $k \in \{0,1\}$ P_k est connu (i.e on a $\hat{P}_k = P_k$). Sous les hypothèses **A0**-d et **A1**, si $\hat{h}$ est la fonction de segmentation donnée par (2.2), alors il existe une constante c_0 positive telle que*

$$\mathcal{S}(\hat{h}) \leq c_0 \left(\frac{\log(N)}{N} \right)^{1/d},$$

où $\mathcal{S}$ est l'excès de risque de segmentation défini par (1.4).

Nous démontrons ce théorème dans la Section 4. Nous donnons dans la section qui suit l'algorithme de Kolaczyk et al. [46] et un résultat que nous avons obtenu pour cet algorithme. Dans la Section 3, nous donnons quelques applications de cet algorithme à des données de géologie sur Mars et aux données médicales.

2.2 Estimation par maximum de vraisemblance pénalisé des poids

Nous ne présentons pas exactement le modèle de mixlet proposé par Kolaczyk et Al. [46] mais une version simplifiée de celui-ci que nous utiliserons. En théorie, puisque nous supposons pour l'instant que les distributions P_k du mélange sont connues, la nature exacte de $\mathcal{X}$ et des distributions $(P_k)_k$ utilisées ne sont pas essentielles. Ce sont les poids du mélange qui sont inconnus.

Afin de différencier les vrais poids du mélange des autres poids que nous envisagerons, nous noterons $\pi^*_{.k} = (\pi^*_{ik})_{i\in\mathcal{T}_N}$ le vecteur des vrais poids inconnus pour la classe k, et $\pi^*_{i.} = (\pi^*_{ik})_{k\in\{1,\dots,K\}}$. Si u_i est une densité sur $\mathcal{X}$ d'une loi du type

$$\sum_{k=1}^{N} \pi_{ik} P_k, \quad \sum_{k=1}^{K} \pi_{ik} = 1, \tag{2.3}$$

nous identifierons u_i et les poids associés $\pi_{.k} = (\pi_{ik})_{k=1,\dots,K}$. Nous noterons s^*_i la densité associée aux vrais poids du mélange, et $s^* = \Pi_{i=1}^N s^*_i$. Ainsi, s^*_i sera la densité de la loi :

$$\sum_{k=1}^{N} \pi^*_{ik} P_k, \quad . \tag{2.4}$$

Chaque poids π_{ik} prend une valeur dans $[0,1]$. Nous nous restreindrons par la suite à la recherche de poids sur une grille de pas $N^{-3/2}$ (ce choix peut être interprété comme la réalisation d'un équilibre entre un échantillonnage trop fin et un échantillonnage pas assez fin. A toute partition $P = \{R_1, \dots, R_l\}$ de $\mathcal{T}_N$ on peut associer la famille de densités produits $u(X) = \Pi_{i=1}^N u_i(X_i)$ sur $\mathcal{X}^N$ construite à partir de densités u_i de lois données par (2.3), et ayant des poids constants sur les régions $\{R_1, \dots, R_l\}$. L'ensemble des partitions récursives dyadiques de $\mathcal{T}_N$ sera noté $\mathcal{P}_N$. Nous noterons $\mathcal{M}_N(\mathcal{P}_N)$ l'union de toutes les densités produits obtenues avec les partitions récursives dyadiques de $\mathcal{T}_N$. $\mathcal{M}_N(\mathcal{P}_N)$, est une famille finie.

Notons que si $u \in \mathcal{M}_N(\mathcal{P}_N)$ il peut exister plusieurs partitions de $\mathcal{P}_N$ auxquelles u est associée. Nous noterons $P(u)$ la plus petite de ces partitions.

Au vu d'un vecteur d'observations $X \in \mathcal{X}^N$, on choisit alors d'estimer la densité s^* par $\hat{s} \in \mathcal{M}_N$ de la manière suivante. La densité $\hat{s}$ choisie réalise parmi $u \in \mathcal{M}_N$ le maximum de vraisemblance pénalisé :

$$\hat{s} = \mathrm{Argmax}_{u\in\mathcal{M}_N} \left\{ \log(u(X)) - 2pen(u) \right\}. \tag{2.5}$$

La fonction $pen(u)$ pénalise les arbres trop riches, elle vaut

$$pen(u) = m^{d-1} \left(\frac{3}{2}(K-1)\log N + \frac{4}{3}\log 2 \right), \tag{2.6}$$

où m^{d-1} est le nombre d'éléments de la partition $P(u) \in \mathcal{P}_N$. La construction de cette pénalité résulte d'un raisonnement simple de la théorie du codage et de l'information. L'Annexe C donne

une petite introduction à la théorie correspondante. L'essentiel est que cette pénalité permet d'obtenir l'inégalité de Kraft (voir Annexe C pour la démonstration) :

$$\sum_{u \in \mathcal{M}_N} e^{-pen(u)} \leq 1. \tag{2.7}$$

Définition 2.1. *Si p et q sont deux densité produits sur l'espace produit $\mathcal{X}^N$, nous appellerons distance moyenne de Hellinger et nous noterons H_N la quantité définie par*

$$H_N(p,q)^2 = \frac{1}{N}\sum_{i=1}^{N} h^2(p_i, q_i), \tag{2.8}$$

où $h(p_i, q_i)$ est la distance de Hellinger définie Chapitre 1 Partie I.

Kolaczyk et Al. [46] donnent le détail de l'algorithme utilisé pour calculer le maximum de vraisemblance défini par (2.5). Cet algorithme hérite de la méthode récursive de construction des partitions récursives dyadiques et a donc une rapidité d'exécution intéressante. Dans le travail de Kolaczyk et Al. [46], $d = 2$. Ils démontrent sous les hypothèses A1 et A0-2 qu'il existe une constante c_0 telle que

$$\mathbb{E}\left[\frac{h^2(s^*, \hat{s})}{N}\right] \leq c_0 \left(\frac{\log N}{N}\right)^{1/2}.$$

Ce résultat vrai peut être facilement amélioré puisque $h^2(s^*, \hat{s}) \leq 2$ et donc $\mathbb{E}\left[\frac{h^2(s^*,\hat{s})}{N}\right] \leq 2/N$. Nous pensons que les auteurs voulaient démontrer le résultat suivant (qui est vrai cf Corollaire 2.1 de cette Section)

$$\mathbb{E}\left[H_N^2(s^*, \hat{s})\right] \leq c_0 \left(\frac{\log N}{N}\right)^{1/2}.$$

Nous ne sommes pas parvenu, avec cette dernière équation, à obtenir une borne satisfaisante dans le problème de segmentation. C'est pour cela qu'il nous a fallu obtenir le théorème plus fort suivant qui est essentiel dans la démonstration que nous donnons (Section 4) du Théorème 2.1.

Theoreme 2.2. *Soit $d \geq 2$. Soient $\hat{s}$ et s^* donnés respectivement par (2.5) et (2.4). Alors, sous les hypothèses **A0**-d et **A1**, il existe une constante positive c_1 telle que*

$$P_{s^*}(NH_N^2(s^*, \hat{s}) \geq \delta) \leq e^{c_1 \log^{1/d}(N) N^{\frac{d-1}{d}} - \delta}. \tag{2.9}$$

Démonstration. La démonstration repose sur le même principe que celui exposé par Birgé dans [14].

La densité $\hat{s}$ estimée par maximum de vraisemblance pénalisée, au vu de l'observation $X = (X_1, \ldots, X_N)$, est aussi un T-estimateur. Plus précisément, $\hat{s}$ a battu tous les autres candidats $v \in \mathcal{M}_N$ au test suivant :

$$\psi(u, v, X) = 1_{\log \frac{u(X)}{v(X)} \geq 2(pen(u) - pen(v))}, \tag{2.10}$$

autrement dit

$$\forall v \in \mathcal{M}_N \;\; \psi(\hat{s}, v, X) = 1. \tag{2.11}$$

Ainsi, nous avons

$$P_{s^*}(NH_N^2(s^*,\hat{s}) \geq \delta) \leq P_s\left(\exists u \in \mathcal{M}_N \ : NH_N^2(s^*,u) \geq \delta \text{ et } \forall v \in \mathcal{M}_N \ \ \psi(u,v,X) = 1\right). \quad (2.12)$$

Nous allons utiliser le lemme suivant.

Lemme 2.1. *Il existe un candidat $v_m^* \in \mathcal{M}_N$ tel que pour tout $u \in \mathcal{M}_N$ on ait*

$$\psi(u,v^*,X) \leq 1_{\log \frac{u(X)}{s^*(X)} \geq 2pen(u) - c_1(m^{d-1}\log(N) + \frac{N}{m} + \frac{1}{\sqrt{N}})},$$

où c_1 est une constante positive.

La démonstration de ce lemme est reportée à la fin de la preuve. Puisque en utilisant (2.12) et la propriété de sous-additivité des probabilités, on a :

$$P_{s^*}(NH_N^2(s^*,\hat{s}) \geq \delta) \leq \sum_{u \in \mathcal{M}_N} P_{s^*}(NH_N^2(s^*,u) \geq \delta \text{ et } \psi(u,v^*,X) = 1),$$

l'équation du Lemme 2.1 implique que

$$P_{s^*}(NH_N^2(s^*,\hat{s}) \geq \delta) \leq$$

$$\sum_{u \in \mathcal{M}_N} P_{s^*}\left(NH_N^2(s^*,u) \geq \delta \text{ et } \log\frac{u(X)}{s^*(X)} \geq 2pen(u) - c_1(m^{d-1}\log(N) + \frac{N}{m} + \frac{1}{\sqrt{N}})\right). \quad (2.13)$$

Notons tout de suite que, en fixant m à $(N/\log(N))^{1/d}$, on a :

$$m^{d-1}\log(N) + \frac{N}{m} + \frac{1}{\sqrt{N}} \leq 3\left(\log(N)\right)^{1/d} N^{\frac{d-1}{d}}.$$

Ce choix de m peut être analysé comme un choix d'équilibre entre biais et variance.

En fixant m à $(N/\log(N))^{1/d}$, on a :

$$P_{s^*}\left(NH_N^2(s^*,u) \geq \delta \text{ et } \log\frac{u(X)}{s^*(X)} \geq 2pen(u) - c_1(m^{d-1}\log(N) + \frac{N}{m} + \frac{1}{\sqrt{N}})\right) \quad (2.14)$$

$$\leq e^{-pen(u)} e^{3c_1(\log(N))^{1/d} N^{\frac{d-1}{d}}} \prod_{i=1}^{N} \mathbb{E}_{s_i^*}\left[\left(\frac{u(X_i)}{s^*(X_i)}\right)^{1/2}\right] \mathbb{1}_{NH_N^2(s^*,u) \geq \delta}$$

(d'après l'inégalité de Markov)

$$\leq e^{-pen(u)} e^{3c_1(\log(N))^{1/d} N^{\frac{d-1}{d}}} e^{-\frac{NH_N^2(s^*,u)}{2}} \mathbb{1}_{NH_N^2(s^*,u) \geq \delta}$$

(Par le même raisonnement que dans la preuve de la proposition 1.5 (Chapitre 1 Partie I : borne de Chernoff). Soit avec (2.13),

$$P_{s^*}(NH_N^2(s^*,\hat{s}) \geq \delta) \leq e^{3c_1(\log(N))^{1/d} N^{\frac{d-1}{d}} - \delta} \sum_{u \in \mathcal{M}_N} e^{-pen(u)}.$$

On termine alors la démonstration en utilisant l'inégalité de Kraft (2.7). □

Démonstration du Lemme 2.1. Le pas de discrétisation des éléments de $\mathcal{M}_N$ est $N^{-3/2}$. Aussi, d'après la Proposition 1.1, il est possible de trouver $v_m^* \in \mathcal{M}_N$ associé à une partition $P(v^*)$ de taille m^{d-1} tel que sur un ensemble de pixels I_+ [1],

$$|\pi_i^* - \pi_i(v_m^*)| \leq N^{-3/2} \tag{2.15}$$

et que le cardinal de $\mathcal{T}_N \setminus I_+$ ne soit pas plus grand que $c\frac{N}{m}$. Notons qu'il existe $C_1 > 0$ tel que

$$pen(v_m) \leq C_1 m^{d-1} \log(N). \tag{2.16}$$

Nous allons séparer les cas $i \in I_+$ et $i \in I_-$.

Si $i \in I_+$, d'après une inégalité de convexité bien connue,

$$\log((v_m^*)_i(X_i)) - \log(s_i^*(X_i))) \leq \frac{(\pi_i^* - \pi_i(v_m^*))(f_1(X_i) - f_2(X_i)))}{s_i^*(X_i)},$$

et donc d'après l'Hypothèse A1,

$$\log\left(\frac{(v_m^*)_i(X_i)}{s_i^*(X_i)}\right) \leq 2B|\pi_i^* - \pi_i(v_m^*)|,$$

ce qui au vu de (2.15) implique :

$$\log\left(\frac{(v_m^*)_i(X_i)}{s_i^*(X_i)}\right) \leq \frac{1}{N}\frac{2B}{N^{1/2}}. \tag{2.17}$$

Si $i \notin I_+$ l'Hypothèse A1 implique

$$\log\left(\frac{(v_m^*)_i(X_i)}{s_i^*(X_i)}\right) \leq \log(B). \tag{2.18}$$

En définitive, les équations (2.17) et (2.18) permettent d'obtenir :

$$\log\left(\frac{v_m^*(X)}{s^*(X)}\right) \leq \frac{2B}{N^{1/2}} + \log(B)\frac{N}{m},$$

ce qui implique :

$$\log\left(\frac{u(X)}{v_m^*(X)}\right) = \log\left(\frac{u(X)}{s^*(X)}\right) + \log\left(\frac{v_m^*(X)}{s^*(X)}\right) \leq \log\left(\frac{u(X)}{s^*(X)}\right) + \frac{2B}{N^{1/2}} + \log(B)\frac{N}{m}.$$

Finalement, cette dernière équation et (2.16) impliquent

$$\log\left(\frac{u(X)}{v_m^*(X)}\right) + 2pen(v_m) \leq$$

$$\log\left(\frac{u(X)}{s^*(X)}\right) + \max(2B, \log(B), 1)\left(\frac{1}{N^{1/2}} + \log(B)\frac{N}{m} + C_1 m^{d-1}\log(N)\right),$$

dont le lemme se déduit directement.

[1] $\pi_i(v_m^*)$ est la proportion du mélange définissant $(v_m^*)_i$

Corollaire 2.1. *Soit $q \geq 1$. Sous les mêmes hypothèses que celles du théorème précédent, il existe $c_0 > 0$ tel que*

$$\mathbb{E}[H_N^{2q}(\hat{s}, s)] \leq c_0 \left(N \log(N)\right)^{-\frac{q}{d}}.$$

Démonstration. Il s'agit exactement de reprendre l'approche de Birgé [14] et sa Proposition 3 :

Lemme 2.2. *Soit Y une variable aléatoire positive telle que*

$$P(Y > y) \leq \alpha e^{-y^2} \quad \textit{pour } y \geq \bar{y} \ \textit{ et } \alpha > 0.$$

Alors, pour tout $q \geq 1$,

$$\mathbb{E}[Y^q] \leq \bar{y}^q \left(1 + \alpha \zeta_q(\bar{y})\right),$$

où ζ_q est une fonction définie sur $\mathbb{R}^+$ décroissante et telle que

$$\forall x \geq cq, \ \ \zeta_q(x) = \frac{q}{2} e^{-x}, \quad \textit{où } c = 1/2 \ \textit{ si } q \leq 2\pi e \quad \textit{et } 0.612 \ \textit{ sinon}.$$

On applique le théorème précédent pour vérifier les hypothèses de ce lemme avec $\bar{y}^2 = (c_1 + 1)\log^{1/d}(N) N^{\frac{d-1}{d}}$, $\alpha = e^{\frac{c_1}{c_1+1}\bar{y}}$, $Y^2 = N H_N^2(s^*, \hat{s})$. On a donc pour $(c_1+1)\log^{1/d}(N) N^{\frac{d-1}{d}} > (cq)^2$

$$\alpha \zeta_q(\bar{y}) \leq \frac{q}{2} e^{\left(-(c_1+1)+\frac{c_1}{c_1+1}\right)\log^{1/d}(N) N^{\frac{d-1}{d}}},$$

ce qui permet d'obtenir le résultat voulu (pour N assez grand et donc, puisque $H_N^2 \leq 2$, pour tout N en modifiant la constante). □

2.3 Application aux données médicales et à l'imagerie satellitaire

2.3.1 Application aux données médicales

Pour le problème médical (dont le contexte est détaillé dans l'introduction de ce mémoire), nous disposons de l'échantillon d'apprentissage donné dans l'application de la partie II (Chapitre 2 Section 4). Nous écartons les Métastases de l'échantillon d'apprentissage, car nos données d'apprentissage ne sont pas encore assez nombreuses pour envisager une séparation des Métastases avec les Glioblastomes (voir les raisons du taux de classification médiocre Partie II Chapitre 2 Section 4). Ainsi, nous avons à notre disposition 62 spectres de quatre groupes différents : 21 Glioblastomes de type A, 9 Glioblastomes de type B, 16 Méningiomes, et 9 tissus sains. Il nous a été donné une image hyperspectrale associée à un Glioblastome mélangeant les deux types (A et B). La segmentation obtenue et celle que l'on devrait obtenir sont données par la Figure 2.2. On peut se rendre compte que la tumeur est assez bien localisée, mais que les différents types de Glioblastomes ne sont pas bien séparés.

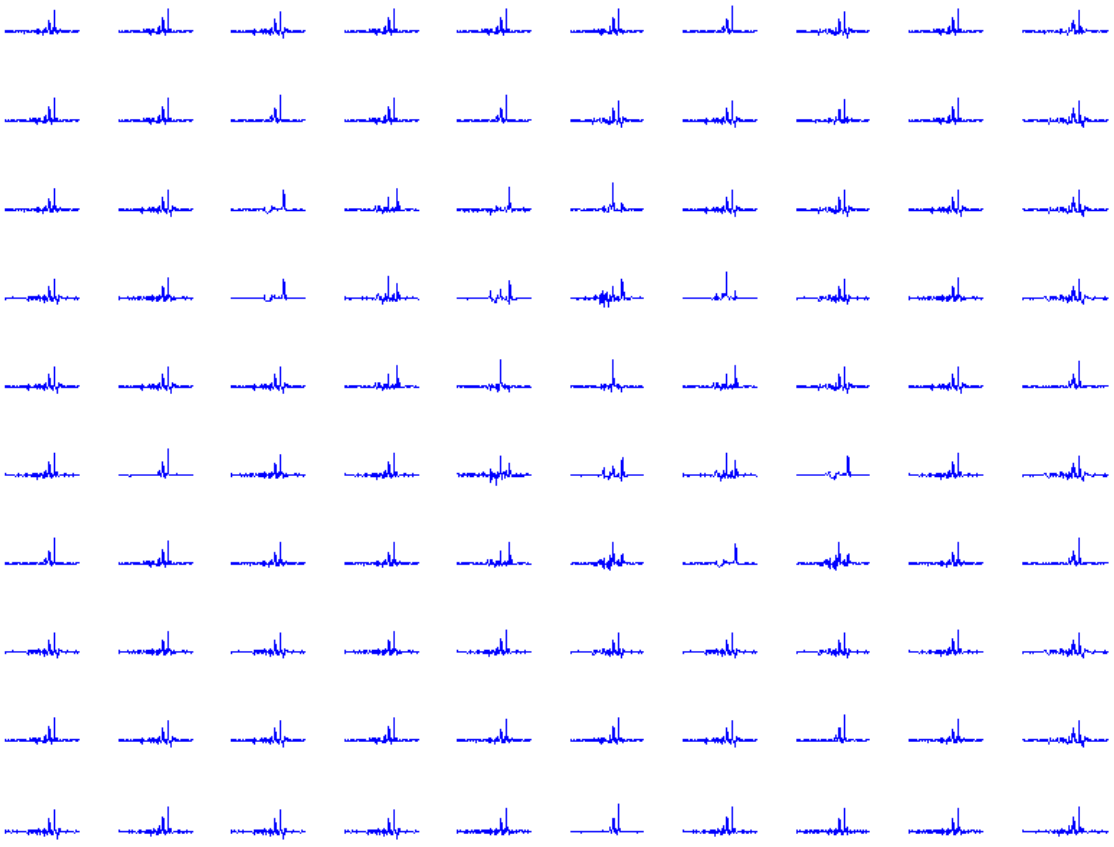


FIG. 2.1 – Carré 10×10 en haut à gauche de l'image Hyperspectrale de Glioblastome de taille 16×16.

Notons que si les résultats sont assez intéressants, ceci résulte d'une part d'un pré-traitement

(rephasage à la main fait par les médecins voir Partie II Chapitre 2 Section 4) assez lourd et du fait que l'on a écarté les Métastases de notre échantillon d'apprentissage. Si les métastases ne sont pas écartées, le résultat est bien moins bon.

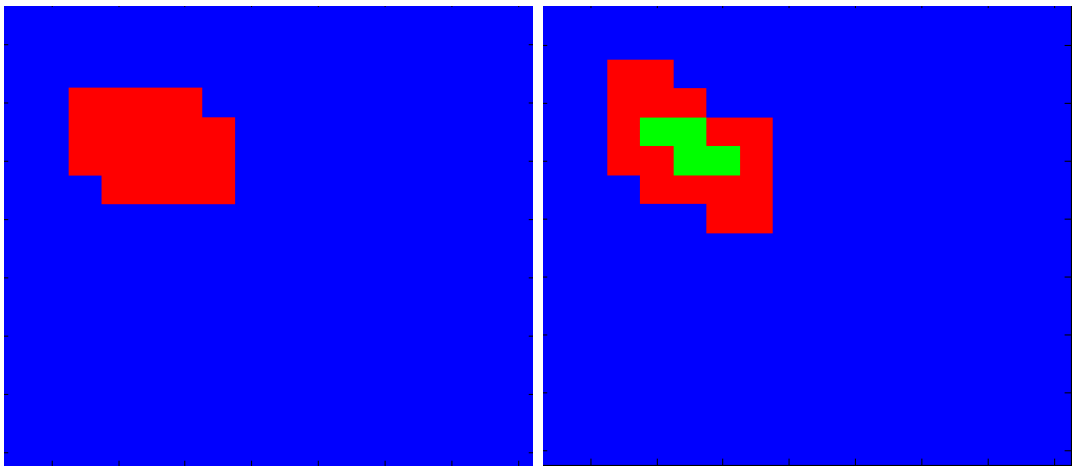


FIG. 2.2 – Segmentation obtenue -à gauche-, et segmentation que l'on devrait obtenir (selon les médecins/physiciens)- à droite-. Les pixels colorés en bleu correspondent à des tissus sain, le vert est du glioblastome de type B et le rouge du glioblastome de type A.

Remarque 2.3 (Sur la réduction de dimension "simultanée")**.** *Dans le cas de la classification (cf Partie II Chapitre 2 Section 2), la réduction de dimension se faisait à partir de l'estimation par seuillage d'un vecteur directeur définissant la frontière entre deux groupes. Ainsi dans le cas de plusieurs classes, nous obtenions plusieurs vecteurs directeurs agissant dans des espaces (en général) différents de dimension différentes. Ici, le problème n'est pas tout à fait le même : on cherche un espace commun à tous les groupes. Ceci est nécessaire au calcul des vraisemblances et peut poser un problème lorsque le nombre de groupes présents devient grand et que les vecteurs directeurs de chacun des hyperplans optimaux séparant les différentes classes sont définis dans des espaces assez distincts. Par exemple, supposons que le nombre de groupes est* $K = 10$*, que la dimension de l'espace des observations est* $p = 256$ *et que pour* $(i, j) \in \{1 \dots, K\}^2$*,* $i \neq j$*, le vecteur directeur de l'hyperplan séparant les groupes* i *et* j *ait cinq composantes non nulles (et significativement grandes). Supposons aussi que ces cinq composantes sont différentes pour deux couples distincts de groupes. La méthode consistant à chercher un espace de dimension réduite, commun à tous les groupes, risque de nous amener à choisir un espace de dimension* $5 * K(K-1)/2 = 245$*, autrement dit quasiment tout l'espace. Ceci constitue le principal défaut de l'algorithme proposé : il oblige l'utilisateur à trouver un espace unique commun à tous les groupes. Notons que l'algorithme présenté dans le Chapitre* 3 *de cette partie n'a pas ce défaut.*

2.3.2 Application à l'imagerie satellitaire

J'ai travaillé durant ma thèse, en collaboration avec Frédéric Schmidt, qui a soutenu le 25 octobre 2007 sa thèse (effectuée sous la direction de Sylvain Douté du laboratoire de planétologie de Grenoble) intitulée : "Classification de la surface de Mars par imagerie hyperspectrale

OMEGA. Suivi spatio-temporel et études des dépôts saisonniers de CO2 et H2O". Les Images de Mars étudiées par celui-ci ont une grande résolution spatiale. Sur chaque pixel de ces images on observe un spectre reflétant la nature physique des matériaux. Chaque spectre est composé de 256 bandes spectrales. Les Physiciens spécialistes de l'instrumentations qui se retrouvent face à ce type de données dans le cadre de leur recherche sont de plus en plus nombreux. Les méthodes utilisées pour observer, analyser ou traiter ces données ne sont pas automatisées.

Nous avons cherché à obtenir une segmentation des images du Laboratoire de Planétologie de Grenoble avec l'algorithme que nous avons donné. Dans l'image étudiée, les types de spectres observés sont issus soit de glace d'eau soit de glace de CO_2 soit de poussière. Nous possédons pour chaque groupe à peu près 2000 spectres discrétisés sur 256 bandes spectrales. Autrement dit le nombre d'observations de la base d'apprentissage est très grand et on peut supposer que la phase d'apprentissage nous a permis de connaître assez bien les différentes distributions associées aux différents groupes. La segmentation obtenue est présentées Figure 2.3. Nous n'avons autorisé les regroupements de pixels qu'à l'échelle la plus fine (les données étant bien séparées, ceci permet d'admettre des frontières assez irrégulière entre les différents groupes). En d'autres termes nous ne tirons pas un grand parti de la régularité des frontières.

Nous avons effectué deux types de traitement. Dans le premier, nous avons appliqué à chaque spectre une transformée de Fourier inverse avant d'appliquer la transformée en ondelette. Dans le deuxième la transformée en ondelette était appliquée directement. La Figure 2.4 donne la décroissance de la statistique S_{RD} (définie à la Partie II Chapitre 2 Section 3 Equation (2.14)) mesurant l'intérêt des différentes dimension pour la classification. Cette décroissance est plus forte dans le cas où une transformée de Fourier inverse a été faite au préalable. Ces résultats présentés Figure 2.3 ont été jugés intéressant par l'équipe des géologues. Ceux-ci n'utilisaient que trois bandes spectrales choisies pour leur intérêt physique et traitait chaque pixel indépendamment des autres. Notre critère de réduction de dimension (et de linéarisation de la règle) nous a amené à nous placer dans un espace de dimension environ 60.

Notre approche apporte donc clairement une sélection automatique des directions spectrales utilisées pour la segmentation (grâce à la méthode décrite au Chapitre 2 de la partie II) et un algorithme rapide de segmentation tenant compte de la régularité des frontières entres les différentes zones de l'image.

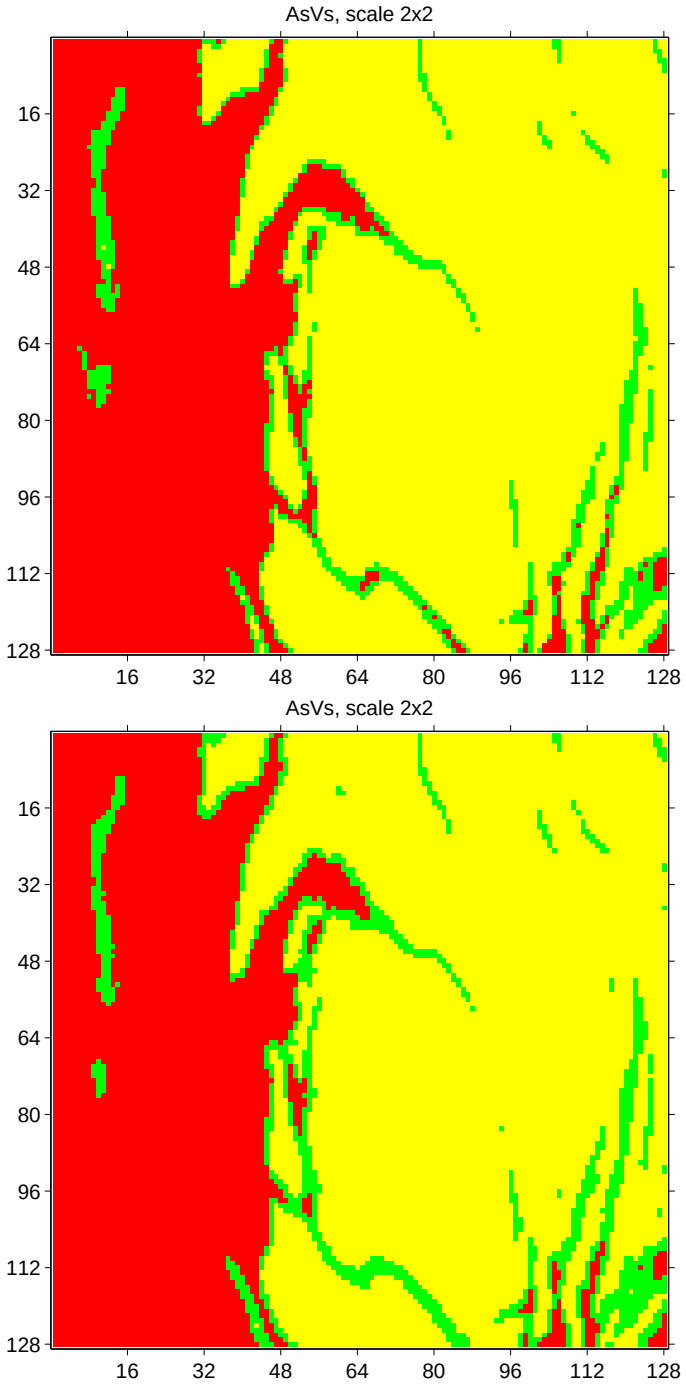


FIG. 2.3 – Segmentation obtenue. Transformée de Fourier inverse puis transformée en ondelette (en haut). Transformée en ondelette (en bas). En rouge poussière, en jaune glace de CO_2 et en vert glace d'eau.

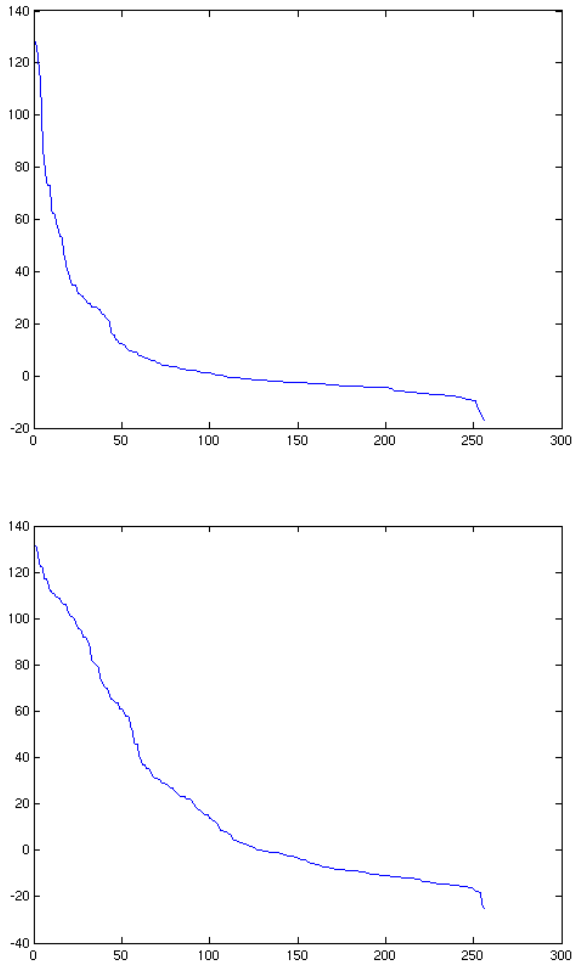


FIG. 2.4 – Décroissance des coefficients de S_{RD} (définie à la Partie II Chapitre 2 Section 3 Equation 2.14). Transformée de Fourier inverse puis transformée en ondelette (en haut). Transformée en ondelette (en bas)

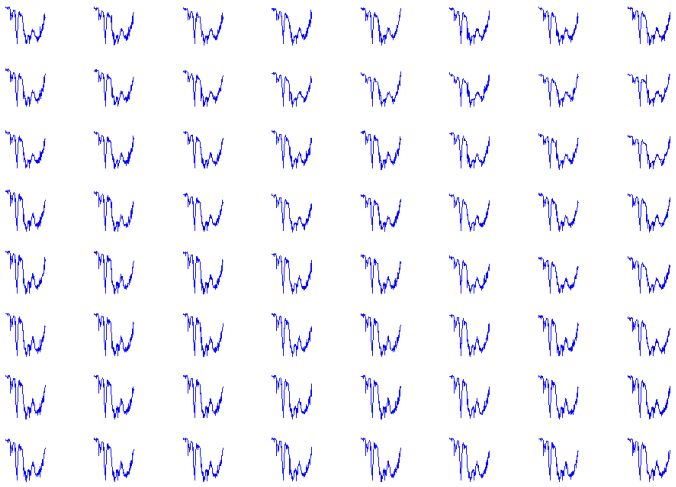


FIG. 2.5 – Portion de l'image hyperspectrale dans la zone $[48:55] \times [120:127]$

2.4 Démonstration du théorème 2.1

Démonstration. Nous allons utiliser dans cette preuve les formulations du problème de learning par plug-in (Chapitre 5 Partie I). Dans le cas où $K = 2$, la règle de Bayes et la règle plug-in s'écrivent :

$$h^*(X_i,i) = \begin{cases} 1 & \text{si } \frac{\pi_{i1} f_1(X_i)}{s_i(X_i)} \leq 1/2 \\ 2 & \text{sinon} \end{cases} \quad \text{et } \hat{h}(X_i,i) = \begin{cases} 1 & \text{si } \frac{\hat{\pi}_{i1} f_1(X_i)}{\hat{s}_i(X_i)} \leq 1/2 \\ 2 & \text{sinon} \end{cases}, \tag{2.19}$$

où f_1 est la densité de P_1 et f_2 la densité de P_2. Dans la suite, nous noterons $\eta_i(X_i) = \frac{\pi_{i1} f_1(X_i)}{s_i(X_i)}$ et $\hat{\eta}_i(X_i) = \frac{\hat{\pi}_{i1} f_1(X_i)}{\hat{s}_i(X_i)}$. Nous remarquons que, d'après le Théorème 5.1 du Chapitre 5 de la Partie I (plus précisément grâce à l'équation (5.7) de ce théorème),

$$\mathbb{E}\left[\sum_{i=1}^{N} 1_{\hat{h}(i,X_i)\neq Y_i} - 1_{h^*(i,X_i)\neq Y_i} | (X_i)_{i=1;\ldots;N}\right] = \sum_{i=1} |2\eta_i(X_i) - 1| 1_{h^*(X_i,i)\neq \hat{h}(X_i,i)}.$$

Aussi nous noterons

$$M = \frac{1}{N}\sum_{i=1} |\eta_i(X_i) - 1/2| 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}, \quad DC = \sum_{i=1}^{N} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)} \text{ et } M^k = M 1_{DC=k}. \tag{2.20}$$

DC est le nombre de pixels différemment classifiés par la règle de Bayes et la règle plug-in. Nous allons utiliser les deux lemmes suivants, démontrés à la suite de cette preuve.

Lemme 2.3. *Il existe $c_0 > 0$ telle que*

$$P\left(M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right] \text{ et } M^k > 0\right) \leq 2e^{-2\frac{k^2}{N}}. \tag{2.21}$$

Lemme 2.4. *Quel que soit $c_3 > 0$, il existe $c_1, c_2 > 0$ tels que*

$$P\left(M^k \geq c_3 \frac{k}{N}\right) \leq e^{c_1 N^{\frac{d-1}{d}} \log^{1/d}(N) - c_2 k}. \tag{2.22}$$

Remarquons que $M \leq 1$. On a $err = E_s[M] = \sum_{k=1}^N E_s[M^k]$ et donc :

$$err \leq \frac{k_0}{N} + \sum_{k=k_0}^{N} \left(P\left(M^k \in \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]\right) + P\left(M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]\right)\right),$$

soit, au vu des deux lemmes :

$$err \leq \frac{k_0}{N} + N\left(2e^{-2\frac{k_0^2}{N}} + e^{c_1 N^{\frac{d-1}{d}} \log^{1/d}(N) - c_2 k_0}\right). \tag{2.23}$$

Le choix $k_0 = \frac{c_1+1}{c_2} N^{\frac{d-1}{d}} \log^{1/d}(N)$ permet de conclure. □

Il nous reste maintenant à démontrer les deux lemmes utilisés.

Démonstration du Lemme 2.3.

Notons que

$$1_{DC=k} = \sum_{I_k} \prod_{i \in I_k} 1_{h^*(X_i,i) \neq \hat{h}(X_i,i)},$$

où la première somme est prise sur toutes les parties I_k de $\{1, \ldots, N\}$ à k éléments. Nous allons poser

$$Z_i = \frac{1}{N}|\eta_i(X_i) - 1/2|,$$

et donc

$$1_{M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]} 1_{M^k > 0} \leq 1_{DC=k} 1_{M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]} = \sum_{I_k} 1_{\sum_{j \in I_k} Z_i \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]} \prod_{i \in I_k} 1_{h^*(X_i,i) \neq \hat{h}(X_i,i)},$$

où M^k est défini par (2.20). Ainsi, on a en prenant l'espérance :

$$P\left(M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right],\ M^k > 0\right) \leq \sum_{I_k} \mathbb{E}\left[1_{\sum_{i \in I_k} Z_i \notin [\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}]} \prod_{i \in I_k} 1_{h^*(X_i,i) \neq \hat{h}(X_i,i)}\right].$$

D'après l'inégalité de Holder, pour tout $q \in [0,1]$

$$P\left(M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right],\ M^k > 0\right) \leq$$

$$\sum_{I_k}\left(\mathbb{E}\left[\prod_{i\in I_k} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}\right]\right)^{1-q} P_s\left(\sum_{i\in I_k} Z_i \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]\right)^q. \tag{2.24}$$

Nous allons borner $P\left(\sum_{i\in I_k} Z_i \notin [\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}]\right)$ en appliquant l'inégalité des différences finies. Nous la rappelons.

Theoreme 2.3 (Inégalité des différences finies)**.** *Soit g une application de $\mathcal{X}^p$ dans $\mathbb{R}$ vérifiant la propriété suivante. Il existe $b_1, \ldots, b_p$ constantes positives telles que*

$$\sup_{x_1,\ldots,x_p,x_i'\in\mathcal{X}} |g(x_1,\ldots,x_p) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_p)| \leq b_i,\ \ 1 \leq i \leq p \tag{2.25}$$

Si $X_1, \ldots, X_p$ sont p variables aléatoires indépendantes à valeur dans $\mathcal{X}$, alors la variable aléatoire $Z = g(X_1, \ldots, X_n)$ vérifie

$$P(|Z - \mathbb{E}[Z]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^p b_i^2}}. \tag{2.26}$$

Le fait que $|\eta_i(X_i) - 1/2|$ soit borné (presque sûrement) supérieurement et d'espérance finie bornée inférieurement par une constante positive nous permet d'obtenir en appliquant l'inégalité des différences finies à $g((Z_i)_{i\in I_k}) = \sum_{i\in I_k} Z_i$, $p = k$, l'existence d'une constante c_k positive, bornée indépendamment de k, telle que $\mathbb{E}[Z] = c_k\frac{k}{N}$. Ce qui donne en fixant $c_0 = hc_k$ et $t = \frac{kc_0}{2N}$ dans (2.26) :

$$P\left(\sum_{i\in I_k} Z_i \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]\right) \leq 2e^{-2k^2}, \tag{2.27}$$

et donc (2.24) deviens, en posant $q = 1/N$:

$$P\left(M^k \notin \left[\frac{k}{N}\frac{c_0}{2}; \frac{k}{N}\frac{3c_0}{2}\right]\right) \leq 2e^{-\frac{2k^2}{N}} \sum_{I_k}\left(\mathbb{E}\left[\prod_{i\in I_k} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}\right]\right)^{1-1/N}. \tag{2.28}$$

D'autre part, d'après l'inégalité de Holder

$$\sum_{I_k}\left(\mathbb{E}\left[\prod_{i\in I_k} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}\right]\right)^{1-1/N} 1^N \leq (\sum_{I_k} 1)^{1/N} \sum_{I_k} \mathbb{E}\left[\prod_{i\in I_k} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}\right],$$

et puisqu'il y a moins de 2^N sous-ensembles de $\{1, \ldots, N\}$ de taille k,

$$\sum_{I_k}\left(\mathbb{E}\left[\prod_{i\in I_k} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}\right]\right)^{1-1/N} \leq 2^{N/N} P(DC = k).$$

Ceci, ainsi que l'inégalité (2.24), impliquent bien l'équation du Lemme 2.21.

Démonstration du Lemme 2.4.

Il s'agit d'appliquer le Théorème 2.2. Rappelons (cf Chapitre 5 Partie I) que si $h^*(X_i,i) \neq \hat{h}(X_i,i)$, alors $|\eta_i(X_i) - 1/2| \leq |\eta_i(X_i) - \hat{\eta}_i(X_i)|$. Par conséquent

$$M^k \leq 1_{DC=k} \frac{1}{N} \sum_{i=1}^{N} |\eta_i(X_i) - \hat{\eta}_i(X_i)| 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}. \tag{2.29}$$

D'autre part, un calcul simple fourni :

$$|\eta_i(X_i) - \hat{\eta}_i(X_i)| \frac{\hat{s}_i(X_i)s_i^*(X_i)}{f_1(X_i)f_2(x_i)} = |\pi_i - \hat{\pi}_i|,$$

et donc, en posant

$$U_i = \frac{f_1(X_i)f_2(X_i)}{|f_1 - f_2|_1 \hat{s}_i(X_i)s_i(X_i)} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)},$$

on a :

$$|\eta_i(X_i) - \hat{\eta}_i(X_i)| 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)} = U_i |f_1 - f_2|_1 |\pi_i - \hat{\pi}_i| = U_i |s_i^* - \hat{s}_i|_1. \tag{2.30}$$

L'hypothèse $A1$ implique qu'il existe une constante c positive telle que

$$U_i \leq c^{1/2} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}, \tag{2.31}$$

et l'inégalité de LeCam (cf Chapitre 1 Partie I) implique que $|s_i^* - \hat{s}_i|_1 \leq h(s_i^*, \hat{s}_i)$. Nous allons noter

$$V_i = c^{1/2} 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)}, \tag{2.32}$$

et nous avons donc, au vu de (2.30) :

$$|\eta_i(X_i) - \hat{\eta}_i(X_i)| 1_{h^*(X_i,i)\neq\hat{h}(X_i,i)} \leq V_i h(s_i^*, \hat{s}_i).$$

En utilisant l'équation (2.29) cela donne alors :

$$M^k \leq 1_{DC=k} \frac{1}{N} \sum_{i=1}^{N} V_i h(s_i^*, \hat{s}_i).$$

Ainsi, d'après l'inégalité de Cauchy-Schwartz,

$$(M^k)^2 \leq 1_{DC=k} \frac{1}{N} \sum_{i=1}^{N} V_i^2 \frac{1}{N} \sum_{i=1}^{N} h^2(s_i^*, \hat{s}_i),$$

et au vu de (2.32) :

$$(M^k)^2 \leq \frac{ckH_N^2(s^*, \hat{s})}{N}.$$

Cette dernière équation implique

$$P\left((M^k)^2 \geq \frac{k^2}{N^2} \left(\frac{3c_0}{2} \right)^2 \right) \leq P\left(NH_N^2(s^*, \hat{s}) \geq k \left(\frac{3c_0}{2\sqrt{c}} \right)^2 \right),$$

et donc avec le Théorème 2.2 l'inégalité (2.22). □

Chapitre 3

AWS fonctionel

La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi.

Albert Einstein

Je devine à travers un murmure,
Le contour subtil des voix anciennes
Et dans les lueurs musiciennes,
Amour pâle, une aurore future !

Paul Verlaine

En dehors d'une section finale contenant certains calculs, ce chapitre est articulé en quatre sections. La première consiste en une description des hypothèses mathématiques et sert d'introduction pour la deuxième dédiée à un algorithme de débruitage d'images hyper-spectrales inspiré de l'algorithme AWS proposé par Polzehl et Spokoiny [59]. Cet algorithme ne produit pas de segmentation et c'est pourquoi, dans la troisième section, nous montrons comment cet algorithme peut fournir un estimateur des contours de régions de l'image. Cet estimateur nous permet d'obtenir une segmentation de l'image. La dernière section est consacrée à des simulations, et notre démarche est comparée à celle de Whitcher et al. [76]. Ces auteurs s'intéressent à des techniques de classification non supervisée fondées sur la décomposition en ondelettes des spectres issus d'une image hyper-spectrale. Ce chapitre a fait l'objet d'une publication [36] dans la revue du signal.

3.1 Généralités et notations

3.1.1 Introduction

L'objectif de ce chapitre est, grâce à une technique statistique de réduction de dimension, de proposer un algorithme de segmentation basé sur une réduction de dimension qui ne soit pas la même en différents points de l'image et particulièrement adapté à la segmentation d'images hyper-spectrales fortement bruitées. La segmentation n'est pas supervisée. Aucune information a priori n'est donnée au statisticien. La qualité principale de cet algorithme par rapport à celui présenté dans le chapitre précédent (outre le fait que celui-ci n'est pas supervisé) est qu'il ne nécessite pas une réduction de dimension commune à tous les spectres.

3.1.2 Formalisme

Nous supposons que l'image étudiée est homogène par régions, c'est-à-dire qu'il s'agit d'une fonction $f : [0,1]^2 \to L^2[0,1]$ (donc à valeur dans un espace de fonctions) pour laquelle il existe M régions distinctes A_m, $m = 1 \dots M$, telles que :

$$\forall x \in [0,1]^2 \quad f(x) = \sum_{m=1}^{M} a_m \mathbb{1}_{A_m}(x). \tag{3.1}$$

Dans l'expression (3.1), a_m désigne un élément de $L^2[0,1]$ et $\mathbb{1}_{A_m}(x)$ est l'indicatrice de l'ensemble A_m. Le nombre M de régions, les régions A_m, et les valeurs a_m de f sur ces régions sont inconnus. On supposera, pour des raisons évidentes d'identifiabilité et sans perte de généralité, que les a_m sont tous distincts. La fonction f est observée sur une grille régulière du plan $X_i = (\frac{i_1}{p}, \frac{i_2}{p})$ avec $i = (i_1, i_2) \in \{1, \dots, p\}^2$ dans un bruit blanc gaussien η_i multiplié par un facteur d'échelle σ_i :

$$S_i = f(X_i) + \sigma_i \eta_i \ ; \ i \in \{1, \dots, p\}^2. \tag{3.2}$$

L'indice i désignera un indice double identifiant de manière unique la position d'un pixel (c'est-à-dire, en terme d'imagerie médicale un voxel), et pour deux pixels différents X_i et X_j, les bruits η_i et η_j ne seront pas corrélés.

Le problème que l'on se pose dans la suite est double. Nous voulons d'une part estimer la fonction f et d'autre part identifier le nombre et les régions sur lesquelles f est homogène (i.e. les zones sur lesquelles f est constante).

Nous rappelons que pour une image composée de niveaux de gris, f est à valeur dans $\mathbb{R}$, pour une image RVB, f est à valeur dans $\mathbb{R}^3$. Dans l'idéal, un spectre bruité S_i (une variable aléatoire à valeur dans $L^2[0,1]$) est observé pour chaque volume élémentaire du plan de coupe du cerveau considéré. En pratique les courbes aléatoires S_i ne sont connues qu'en un ensemble de points discret. Elles devront donc être approchées par certaines fonctions définies sur tout $[0,1]$, et cette approximation sera, dans notre cas, réalisée en développant les S_i dans une base d'ondelettes, les coefficients de la décomposition étant estimés à partir des données discrétisées. Pour notre application, nous utiliserons une transformée en ondelettes périodiques (voir par exemple [53]) et nous observons donc en chaque pixel i, $N = 2^J$ coefficients d'ondelette bruités,

notés $\theta_\nu(X_i)$:

$$(P_{E_N}(S_i))_\nu = Y_\nu(X_i) = \theta_\nu(X_i) + \sigma(X_i)\epsilon_\nu(X_i) \quad \nu \in \{1 \dots N\} \tag{3.3}$$

où P_{E_N} désigne la projection du spectre S_i sur l'espace d'approximation d'échelle J. Grâce à l'orthogonalité de la transformée en ondelette, les $\epsilon_\nu(X_i)$ constituent encore un bruit blanc gaussien. L'indice $\nu = (j,k)$ est un couple composé d'un indice d'échelle (noté $|\nu| = j$) et d'un indice de position k.

Il est connu que pour une grande classe de fonctions, la représentation dans le domaine des ondelettes est creuse (voir Annexe A), autrement dit la plupart des fonctions usuelles sont caractérisées par un petit nombre de coefficients non nuls dans la base d'ondelette et ceci sera exploité dans la suite. Toutes nos estimations seront faites sur les coefficients d'ondelette. Il est alors possible d'estimer $f(X_i)$ par une transformée inverse des coefficients d'ondelette estimés.

3.1.3 Synthèse de l'algorithme

Avant de présenter en détail les trois phases différentes qui vont nous permettre d'aboutir à la segmentation d'une image hyper-spectrale, nous en donnons un aperçu rapide.

- Phase 1 : Estimation de $w_{ij} = 1_{f(i)=f(j)}$ par $\hat{w}_{ij}$ grâce à la combinaison d'un procédé de réduction de dimension et de l'algorithme AWS de Pozhel et Spokoiny [59] (Section 3.2).
- Phase 2 : Détection des frontières par un système de votes utilisant l'estimation des poids fournie par la phase 1 (Section 3.3).
- Phase 3 : Regroupement des régions obtenues lors de la phase 2 par une méthode de minimisation de l'erreur quadratique empirique pénalisée (Section 3.4).

3.2 AWS : Algorithme de débruitage et d'estimation des poids

L'algorithme que nous allons décrire et utiliser est inspiré dans ses grandes lignes par l'algorithme AWS de Pozhel et Spokoiny [59]. Ces derniers ont appliqué leur algorithme à des données vectorielles [60] (des données modélisées par des fonctions du type (3.1) pour lesquelles a_m est un vecteur réel multidimensionnel). Leur application est donc radicalement différente de celle que l'on considère ici, car dans leur démarche, les vecteurs a_m ne tiennent pas compte de l'aspect fonctionnel des données. En d'autres termes, l'originalité de la méthode décrite ici réside dans la combinaison d'une méthode de réduction de dimension (qui correspond à utiliser le caractère fonctionnel des données, ou de la grande dimension des données (cf Chapitre 3 partie I)) et de l'algorithme AWS.

L'algorithme est itératif. Etant donné une estimation $(\hat{\theta}_i^{k-1})_i$ des vecteurs des coefficients d'ondelette de f obtenue à l'étape $k-1$, pour chaque pixel i, la k^{eme} itération est composée de deux étapes :

- étape 1 : Sélectionner le sous ensemble V_i^{k+} des pixels de V_i^k voisins de i qui appartiennent à la même zone que i grâce à $(\hat{\theta}_i^{k-1})_i$ (la méthode de sélection correspondante est décrite au paragraphe suivant).
- étape 2 : Estimer la valeur de θ_i le vecteur des coefficients d'ondelette au pixel i, grâce à une moyenne des observations sélectionnées à l'étape 1.

La suite de voisinages $(V_i^k)_k$ des pixels susceptibles d'appartenir à la même zone que i est destinée à croître avec k et doit être choisie par l'utilisateur de l'algorithme. L'important est que dans les premières étapes de l'algorithme, le voisinage soit suffisamment petit pour permettre une réduction du bruit sans trop d'erreurs de sélection.

Après une description de la méthode de sélection, l'algorithme sera détaillé de manière plus formelle et plus précise dans la Section 3.2.3.

Remarque 3.1. *Pozhel et Spokoiny [59] proposent une troisième étape dans leur algorithme. Pour les résultats théoriques qu'ils obtiennent (qui ne s'appliquent pas correctement à notre cas de figure), cette étape, basée sur l'algorithme de Lepski (cf [51]), est fondamentale. Dans la pratique, nous n'avons pas noté de différence significative lorsque cette étape était supprimée. Les preuves données par Pozhel et Spokoiny (voir l'article théorique [61]) sont basées sur un certain nombre d'hypothèses, certaines fausses. Dans notre cas, on peut penser que l'origine du bon fonctionnement de AWS n'est pas dans l'utilisation de cette troisième étape.*

3.2.1 Méthode de sélection

Pour chaque pixel i, l'algorithme que nous allons décrire estime progressivement l'unique zone $A_{m(i)}$ à laquelle i appartient. Il sélectionne les pixels d'un voisinage V_i de i qui sont vraisemblablement dans $A_{m(i)}$. Cette sélection est formalisée grâce à un ensemble de tests des hypothèses suivantes :

$$\forall j \in V_i, \quad H_{0j} : j \in A_{m(i)} \ \text{ contre } \ H_{1j} : j \notin A_{m(i)}. \tag{3.4}$$

L'estimateur de $V_i^+ = V_i \cap A_{m(i)}$ est tout naturellement :

$$\widehat{V_i^+} = \{ j \in V_i \mid H_{0j} \text{ accepté } \}. \tag{3.5}$$

Il s'agit donc à présent de tester les hypothèses de (3.4), ce qui revient à tester si les fonctions de $L^2[0,1]$, $f(X_i)$ et $f(X_j)$ sont identiques ou pas pour $j \in V_i$, i.e :

$$\forall j \in V_i, \quad H_{0j} : \|f(X_j) - f(X_i)\|_{L^2[0,1]} = 0 \ \text{ contre } \ H_{1j} : \|f(X_j) - f(X_i)\|_{L^2[0,1]} \neq 0. \tag{3.6}$$

Le test que nous utilisons et que rappelons a été introduit dans [68] et [30]. C'est le test sur la norme l^2 avec seuillage présenté dans le Chapitre 3 de la Partie I. Remarquons que l'orthogonalité de la base d'ondelettes périodiques de $L^2[0,1]$ permet de réécrire (3.6) de la manière suivante :

$$\forall j \in V_i, \quad H_{0j} : \|(\theta_\nu(X_j))_{\nu\in\mathbb{N}} - (\theta_\nu(X_j))_{\nu\in\mathbb{N}}\|_{l^2} = 0 \ \text{ contre } \ H_{1j} : \|(\theta_\nu(X_j))_{\nu\in\mathbb{N}} - (\theta_\nu(X_j))_{\nu\in\mathbb{N}}\|_{l^2} \neq 0. \tag{3.7}$$

Au vu de l'approximation de f par sa projection dans E_N nous confondrons les hypothèses de (3.7) avec :

$$\forall j \in V_i, \quad H_{0j} : \|(\theta_\nu(X_j))_{\nu\in\{1\dots N\}} - (\theta_\nu(X_j))_{\nu\in\{1\dots N\}}\|_2 = 0$$

$$\text{contre } \ H_{1j} : \|(\theta_\nu(X_j))_{\nu\in\{1\dots N\}} - (\theta_\nu(X_j))_{\nu\in\{1\dots N\}}\|_2 \neq 0$$

ce qui revient finalement à tester la nullité de la moyenne d'un certain nombre de vecteurs gaussiens de dimension N avec N grand.

3.2.2 Test d'hypothèses fonctionnelles : réduction de dimension

Soit μ un vecteur de $\mathbb{R}^N$ composé des $N = 2^J$ premiers coefficients d'ondelette d'une fonction g. Nous supposons que ce vecteur est observé dans un bruit gaussien à composantes indépendantes et identiquement distribuées de variance σ et de moyenne nulle :

$$Z_\nu = \mu_\nu + \sigma\epsilon_\nu \quad \nu = 1 \dots N.$$

Nous allons décrire une procédure pour tester, au vu de ces observations, les hypothèses :

$$H_0 : \|\mu\|_2 = 0 \quad \text{Vs} \quad H_1 : \|\mu\|_2 > \rho_N. \tag{3.8}$$

La procédure correspondante est celle décrite au Chapitre 3 de la Partie I. Nous supposons σ connu. En pratique, nous avons utilisé l'estimation définie par la médiane des valeurs absolues des écarts à la médiane divisée par $0,6745$, appliquée aux coefficients d'ondelette à l'échelle la plus fine. Ceci produit un estimateur robuste de σ (voir par exemple [53] au Chapitre X).
Il est nécessaire lorsque l'on veut tester la nullité de μ dans (3.8) que l'hypothèse alternative soit suffisamment séparée de l'hypothèse nulle (voir Chapitre 2 de la Partie I) et ce, afin que le test n'ait pas une puissance trop faible. La valeur ρ_N est un seuil de séparabilité entre l'hypothèse nulle et l'alternative garantissant une puissance satisfaisante.

Nous rappellons que pour une vaste gamme de fonctions g, μ est creux, c'est-à-dire ne contient que peu de coefficients non nuls (ou d'amplitude significative). Cela explique que la construction de la statistique de test passe par l'estimation aux niveaux j tels que $j_s \leq j \leq \log_2(N)$ des ensembles $I_j \subset \{\nu \text{ tels que } |\nu| = j\}$ des indices des coefficients d'ondelette non nuls. Les coefficients d'échelles plus grossières sont conservés :

$$\widehat{I}_j = \begin{cases} \left\{\nu \text{ tels que } |\nu| = j \quad \text{et} \quad |Z_\nu| > \sigma 4\sqrt{8\log(2^{j-j_s})}\right\} & j_s \leq j \leq \log_2(N) = J \\ \{\nu \text{ tels que } |\nu| = j\} & 0 \leq j \leq j_s \end{cases}. \tag{3.9}$$

L'énergie de ces coefficients est donnée par :

$$\mathcal{E}(Z)^2 = \sum_{\nu \in \cup_{j=1}^J \widehat{I}_j} Z_\nu^2. \tag{3.10}$$

Si cette énergie est trop importante, on choisit de rejeter H_0. La construction de la statistique de test correspondante, ainsi que le choix du seuil à partir duquel on décide de rejeter H_0, nécessitent le calcul de $\mathbb{E}[\mathcal{E}(Z)^2]$ et de $\mathrm{Var}(\mathcal{E}(Z)^2)$ sous H_0 (notés $\mathbb{E}_0[\mathcal{E}(Z)^2]$ et $\mathrm{Var}_0(\mathcal{E}(Z)^2)$). Un calcul exact de ces expressions est donné dans [1]. La statistique de test est obtenue en centrant et en normalisant l'énergie :

$$T(Z) = \frac{\mathcal{E}(Z)^2 - \mathbb{E}_0[\mathcal{E}(Z)^2]}{\sqrt{\mathrm{Var}_0(\mathcal{E}(Z)^2)}}. \tag{3.11}$$

Sous H_0, cette statistique suit asymptotiquement une loi normale centrée réduite (voir [68]). La région de rejet correspondante est :

$$R_\lambda^H = \{T > \lambda\}$$

où λ peut être choisi pour assurer une probabilité de fausse alarme plus petite que α : $\lambda = z_{1-\alpha}$ (z_α est le quantile d'ordre α de la loi normale centrée réduite) . Nous renvoyons à ce sujet le lecteur au Chapitre 3 de la Partie I et aux remarques qui y sont faites.

Remarque 3.2. *La recherche du paramètre j_s (le nombre de niveaux qui ne sont pas seuillés) doit en théorie être guidée par la connaissance a priori de la régularité de la fonction. Spokoiny ([68]) propose une méthode de sélection de j_s par un algorithme adaptatif. Cependant, dans la pratique avec des tailles de trajectoires de 256 et pour notre application en particulier, le choix $j_s = 3$ réalise un bon compromis.*

3.2.3 L'algorithme AWS

A ce stade et par la suite, nous noterons $\theta_i = (\theta_\nu(X_i))_{\nu\in\{1\dots N\}}$, et $Y_i = (Y_\nu(X_i))_{\nu\in\{1\dots N\}}$, le vecteur des coefficients d'ondelette à estimer au pixel i et les observations correspondantes introduites en (3.3).

Rappelons que si V_i est un voisinage du pixel i, nous avons noté V_i^+ l'ensemble des pixels de V_i qui sont dans la même zone que le pixel i. Pour chaque pixel i on définit une suite croissante de voisinages $(V_i^k)_k$. Le plus petit d'entre eux, point de départ de l'algorithme, est réduit à un pixel : $V_i^0 = \{i\}$. Il est évident que $V_i^{0+} = A_{m(i)} \cap V_i^0 = \{i\}$. Etant donné une estimation $\widehat{V_i^{k+}}$ de l'ensemble $V_i^{k+} = A_{m(i)} \cap V_i^k$ des pixels du voisinage de i qui sont dans la même zone que i, un bon estimateur de θ_i est simplement obtenu par la moyenne des observations de $\widehat{V_i^{k+}}$:

$$\hat\theta_i^k = \frac{1}{|\widehat{V_i^{k+}}|} \sum_{j\in\widehat{V_i^{k+}}} Y_i,$$

$|\widehat{V_i^{k+}}|$ désignant le cardinal de $\widehat{V_i^{k+}}$.

L'algorithme consiste à itérer l'étape de sélection suivante de $k = 1$ à k_{max} :

Sélection : Pour chaque pixel i, la sélection se décompose en 2 étapes :

1. Pour chaque pixel j de V_i^k, poser :

$$\Delta_{ij}^k = \hat\theta_i^{k-1} - \hat\theta_j^{k-1}. \tag{3.12}$$

Appliquer la technique de réduction de dimension décrite dans la Sous-section 3.2.2 :

$$\forall j \in V_i^k, \quad T_{ij}^k = T(\Delta_{ij}^k), \ T \text{ étant donné par (3.11).}$$

2. D'après (3.5), estimer V_i^{k+} par l'ensemble des j pour lesquels H_{0j} est accepté :

$$\widehat{V_i^{k+}} = \left\{ j \in V_i^k \text{ tels que } T_{ij}^k \le \lambda_i^k \right\}.$$

L'estimateur final de θ_i est donc un estimateur à noyau :

$$\hat\theta_i^{k_{max}} = \frac{1}{|\widehat{V_i^{k_{max}+}}|} \sum_{j\in\widehat{V_i^{k\max+}}} Y_j = \sum_j \frac{\hat w_{ij}^{k_{max}}}{|\widehat{V_i^{k_{max}+}}|} Y_j \quad \text{avec } \hat w_{ij}^{k_{max}} = 1_{\widehat{V_i^{k_{max}+}}}. \tag{3.13}$$

Remarque 3.3. *Si à l'étape $k-1$ un certain nombre de sélections à tort ont été faites, Δ_{ij}^k a une grande probabilité de ne pas être d'espérance nulle sous H_{0j}. Ceci peut être la cause de nouvelles erreurs et on peut craindre que ce biais ne se propage d'étape en étape. D'autre part, si on cherche à diminuer λ (pour éviter les erreurs), peu de pixels seront sélectionnés ; la variabilité de la statistique de test sera donc trop importante et induira une trop forte probabilité d'erreur : faux positifs (sélection à tort) et faux négatifs (non sélection à tort). Tout le travail pratique et théorique consiste donc à trouver la bonne valeur de λ et à vérifier qu'asymptotiquement pour une certaine classe de valeurs « admissibles » de λ, le biais d'estimation n'explose pas et la variabilité de l'estimation décroît vers* 0. *Néanmoins, il est difficile d'obtenir des résultats théoriques sur la convergence de cette procédure et ce à cause de (3.13) et de la manière dont sont construits les poids. Cette construction établit en effet une structure de dépendance très complexe entre les poids d'une part et entre les poids et les observations d'autre part.*

D'après cette remarque, le choix de λ_i^k est crucial, il conditionne le bon comportement de l'algorithme. Cependant par soucis de clarté, les deux méthodes de sélection, celle de Pozhel et Spokoiny et une autre procédure que nous avons envisagée, sont exposées dans la Section 3.7.

Finalement, l'algorithme AWS est une méthode d'estimation de f. L'estimateur final de f au pixel i étant obtenu par une moyenne des observations dans une fenêtre sélectionnée $\hat{V}_i^{k+}$, elle s'apparente à une méthode d'estimation à noyau non linéaire. Elle ne produit pas de segmentation de l'image, cependant les $\hat{w}_{ij}$ sont une estimation des poids qui seront utilisés dans la section suivante pour la phase de segmentation :

$$w_{ij} = 1_{f(i)=f(j)} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ sont dans la même zone,} \\ 0 & \text{sinon .} \end{cases} \tag{3.14}$$

L'utilisation de ces poids fournis par l'algorithme AWS a déjà été envisagée par Philippe et al. [57] pour détecter des ruptures dans un signal unidimensionnel mais ne l'a jamais été, à notre connaissance, pour la segmentation.

Remarque 3.4. *Si à chaque étape k de l'algorithme en élargissant V_i^k on cherche à faire grandir l'ensemble des pixels j dont on pense qu'ils sont dans la même zone que i, cette méthode n'est pas une variante des méthodes de segmentation par croissance de régions telles qu'elles sont décrites par exemple dans l'article de Zhu et Yuille [78]. Les régions ne sont pas en « compétition » : pour chaque pixel i un ensemble de poids $\hat{w}_{ij}$ est obtenu. Cependant, il n'est pas exclu que $\hat{w}_{ij} = 1$, $\hat{w}_{ik} = 1$ et que $\hat{w}_{jk} = 0$.*

3.3 Segmentation par estimation des frontières

Nous allons présenter ici notre méthode de segmentation. Cette méthode repose sur l'extraction des contours par un système de votes. L'originalité de la méthode réside dans l'exploitation statistique (basée sur le système de votes) des poids de AWS pour la segmentation. Une méthode simple est ensuite utilisée pour obtenir les régions de l'image délimitées par les contours obtenus.

3.3.1 Estimation de frontières par une méthode de vote

Dans le but d'obtenir une description « inter-pixellaire » des frontières (voir Figure 3.1), nous allons chercher à construire une statistique qui indiquera de manière fiable si un segment séparant deux pixels adjacents constitue une frontière ou non. A ce stade et par la suite, $\mathcal{S}$ désignera l'ensemble des segments séparant deux pixels adjacents dans l'image. Pour une image carrée, composée de p^2 pixels, le cardinal de $\mathcal{S}$ est $2p(p-1)$. A un élément u de $\mathcal{S}$ sont associés les deux pixels qu'il sépare : $u_g \in \{1,\dots,p\}^2$ et $u_d \in \{1,\dots,p\}^2$. Un tel segment est une frontière entre deux régions si et seulement si $f(u_g) \neq f(u_d)$. Il est donc possible, comme cela a été fait par Philippe *et al.* [57], pour détecter des ruptures dans un signal, de décider que le segment u sépare les deux pixels adjacents u_g et u_d si $\hat{w}_{u_d u_g} = 0$.

Néanmoins, cette estimation des segments de frontière est trop variable. Pour réduire cette variabilité, il est possible, d'utiliser plus d'informations. En effet, un segment u de $\mathcal{S}$ est un morceau de frontière si et seulement si

$$\forall i \in \{1,\dots,p\}^2 \quad 1_{f(i)\neq f(u_g)} \neq 1_{f(i)\neq f(u_d)}. \tag{3.15}$$

Cette caractérisation d'un segment de frontière utilise les observations associées à tous les pixels. Nous appellerons vote du pixel i pour la frontière u entre deux pixels voisins u_g et u_d, la variable aléatoire :

$$Y_i(u) = 1_{\hat{w}^k_{iu_g} \neq \hat{w}^k_{iu_d}}.$$

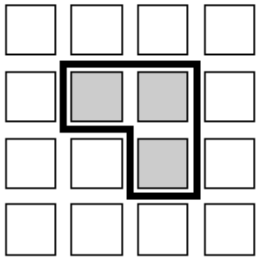


FIG. 3.1 – Description inter-pixellaire d'une frontière

Plus un morceau de frontière totalise de votes, plus il est probable qu'il soit en effet une frontière. La totalité des votes $Y_i(u)$ comptabilisés pour des votants i contenus dans un voisinage $Vois(u)$ sera notée $M(u, Vois(u))$. A un ensemble de segment Γ on peut donc associer la statistique qui comptabilise le nombre normalisé de votes pour les segments de Γ :

$$M(\Gamma) = \frac{1}{\sqrt{n(\Gamma)}} \sum_{u\in\Gamma} M(u, Vois(u)) \;\text{ où }\; M(u, Vois(u)) = \sum_{i\in Vois(u)} Y_i(u) \;\text{ et }\; n(\Gamma) = \sum_{u\in\gamma} |Vois(u)|.$$

En admettant que tous les poids sont indépendants (ce qui n'est qu'une approximation) l'inégalité de Hoeffding permet d'obtenir la proposition suivante :

Proposition 3.1. *Si un ensemble de segments inter-pixelaire Γ n'est composé d'aucune frontière, et que les variables aléatoires Y_i sont indépendantes identiquement distribuées, alors*

$$P_0\left(M(\Gamma) - \mathbb{E}_0[M(\Gamma)] > u_\delta\right) \leq \delta.$$

pour un seuil $u_\delta = \left(\frac{1}{2}\log(1/\delta)\right)^{1/2}$ (l'espérance $\mathbb{E}_0[\,]$ étant prise sous l'hypothèse qu' il n'y a pas de frontières dans Γ).

Le calcul de $\mathbb{E}_0[M(\Gamma)]$ ainsi qu'une méthode pour l'estimer est donnée dans la Section 3.7. Cette proposition donne une valeur de seuil à partir de laquelle un ensemble de segments Γ est considéré comme une frontière. Nous déciderons qu'à un ensemble de segments Γ cohérents (voir sous section suivante) correspondra une frontière si

$$M(\Gamma) - \widehat{\mathbb{E}_0[M(\Gamma)]} > u_\delta, \tag{3.16}$$

où $\widehat{\mathbb{E}_0[M(\Gamma)]}$ est une estimation $\mathbb{E}_0[M(\Gamma)]$ donnée en annexe. Dans les applications, nous avons utilisé $\delta = 0,01$.

3.3.2 Partition de l'image en zones homogènes

Une zone connexe A_i qui n'est pas en contact avec les bords de l'image est délimitée par une frontière $\partial A_i = U_l \partial A_i^l$ où chaque morceau de frontière ∂A_i^l est constitué de segments de $\mathcal{S}$ et forme un cycle sans boucle de $\mathcal{S}$. Si la zone connexe A_i est en contact avec le bord de l'image, alors sa frontière est composée de $\partial A_i = U_l \partial A_i^l$ où chaque ∂A_i^l est constituée de segments de $\mathcal{S}$ et forme soit un cycle sans boucle de $\mathcal{S}$, soit relie un point du bord de l'image à un autre.

Nous allons donc chercher dans une zone $\mathcal{P}$ de l'image un cycle ou une suite de segments reliant un bord de $\mathcal{P}$ à un autre bord de $\mathcal{P}$. La procédure correspondante est appelée recherche-une-frontière($\mathcal{P}$). Cette procédure retourne une frontière Γ dont la somme des poids est aussi importante que possible. Trois cas sont alors envisageables :

- Soit Γ vérifie (3.16) et divise $\mathcal{P}$ en deux régions. Dans ce cas, la frontière est acceptée et la procédure est rappelée pour ces deux régions de l'image.
- Soit Γ vérifie (3.16) et ne divise pas $\mathcal{P}$ en deux régions. Dans ce cas, la frontière est acceptée et la procédure est de nouveau appliquée à l'ensemble $\mathcal{P}$ et en considérant Γ comme une frontière.
- Si Γ ne vérifie pas (3.16) alors la procédure s'arrête, et Γ n'est pas considéré comme une frontière.

En appliquant cette procédure récursive à l'image toute entière, on produit donc une segmentation de cette dernière : on divise l'image en $B_1, \ldots, B_n$ ensembles disjoints de pixels.

Dans notre application, nous avons mis en place une procédure recherche-une-frontière($\mathcal{P}$) qui choisit le segment u tel que $M^k(u)$ soit maximum dans $\mathcal{P}$. Ensuite, on se déplace vers le segment v en contact avec u qui maximise $M^k(v)$. Cette démarche est répétée jusqu'à l'obtention d'un cycle sans boucle ou d'une suite de segments reliant un bord de $\mathcal{P}$ à un autre bord de $\mathcal{P}$ ou d'une suite de segments se recoupant.

3.4 Regroupement des zones

L'étape précédente a permis d'obtenir $B_1, \ldots, B_n$, n ensembles disjoints de pixels (que l'on assimile à des indices doubles) sur lesquels f est supposée être constante. Notre but est de chercher à regrouper les composantes non nécessairement connexes de l'image que la procédure précédente n'a pas regroupées. Soit $\mathcal{P}$ l'ensemble des partitions de l'image que l'on peut obtenir en fusionnant certains des $(B_i)_{i=1,\ldots,n}$. Si p est un élément d'une partition $P \in \mathcal{P}$, $Ave(p)$ désignera la moyenne des observations de p. On cherchera à garder la partition $\hat{P}$ qui rend petit le risque quadratique :

$$\mathcal{R}_P = \sum_{p \in P} \sum_{i \in p} \frac{\|Y_i - Ave(p)\|_{2,n}^2}{\sigma(X_i)^2},$$

où $\|x\|_{2,n} = \frac{1}{n} \sum_{i=1}^n x_i^2$ et $\sigma(X_i)$ est calculé selon la procédure décrite dans la Section 3.2.2.

Avec ce critère, il est clair que c'est la partition qui possède le plus de régions qui va être favorisée. Aussi est-il nécessaire d'appliquer une pénalité, fonction du nombre de régions. Cette pénalité est le plus souvent assez complexe, mais une pénalité proportionnelle au nombre d'éléments de la partition (autrement dit à la dimension du modèle, qui correspond à supposer que notre fonction f est constante sur chaque partie de la partition). La pénalité $|P|\eta$ permet d'obtenir de bons résultats autant d'un point de vue théorique que pratique. La segmentation finale choisie est donc

$$\hat{P} = \text{Argmin}_{P\in\mathcal{P}}\{\mathcal{R}_P + |P|\eta\}.$$

Le paramètre η peut être choisi par un critère de validation croisé, mais un autre choix raisonnable est également $\eta = 2log(p^2)$ (voir par exemple [24] ou [45] pour une justification théorique).

D'un point de vu algorithmique, nous utiliserons le principe de CART ([16]). Etant donnée une partition $P \in \mathcal{P}$ composée de n_P régions, on choisit de regrouper les deux éléments p_1 et p_2 de cette partition qui minimisent

$$\mathcal{R}_{\{p_1\cup p_2\}} = \sum_{i\in p_1\cup p_2} \frac{\|Y_i - Ave(p_1\cup p_2)\|_2^2}{\sigma(X_i)^2}.$$

Si la nouvelle partition $\tilde{P}$ induite par un tel regroupement vérifie :

$$\mathcal{R}_{\tilde{P}} - \mathcal{R}_P < 2\log(p^2),$$

c'est-à-dire si

$$\mathcal{R}_{\{p_1\cup p_2\}} - (\mathcal{R}_{\{p_1\}} + \mathcal{R}_{\{p_2\}}) < 2\log(p^2),$$

alors le regroupement est accepté. On cherche alors un nouveau regroupement. Sinon il est rejeté et l'algorithme s'arrête. L'astuce algorithmique consiste à calculer $\mathcal{R}_{\{p_1\cup p_2\}}$ grâce à la formule de Huygens :

$$\mathcal{R}_{\{p_1\cup p_2\}} = \mathcal{R}_{\{p_1\}} + \mathcal{R}_{\{p_2\}} + |p_1|(Ave(p_1) - Ave(p_1\cup p_2))^2 + |p_2|(Ave(p_2) - Ave(p_1\cup p_2))^2,$$

$$Ave(p_1\cup p_2) = \frac{1}{|p_1\cup p_2|}\left(|p_1|Ave(p_1) + |p_2|Ave(z_2)\right).$$

3.5 Application à des données médicales.

Nous présentons maintenant les résultats obtenus par notre algorithme pour une expérience proposée par Witcher et al. [76]. L'image hyper-spectrale est une image 64×64, pour laquelle en chaque pixel, un spectre discrétisé sur 128 points est observé dans un bruit blanc gaussien. L'image est divisée en 12 régions actives (celles sur lesquelles le signal est non nul) réparties en trois colonnes de quatre carrés de 16×16 pixels. Sur chacun des pixels de ces carrés le signal est de la forme $f(t) = a + b(e^{-t/T_{out}} - e^{-t/T_{in}})$ (voir Figure 3.2). Le lecteur pourra consulter [76] pour une explication plus détaillée. Ces signaux temporels sont une version simplifiée de ce que les équations théoriques de Bloch (équation différentielle de conservation du moment magnétique) permettent d'obtenir lors d'une séquence d'acquisition RARE (ces signaux ne sont donc pas observés dans le domaine fréquentiel). Les paramètres T_{out} et T_{in} changent d'une colonne à l'autre mais restent les mêmes d'une ligne à l'autre. Les amplitudes maximales de $f(t)$ varient

d'une ligne à l'autre et restent les mêmes d'une colonne à l'autre afin que le rapport contraste sur bruit (CNR) dans chaque ligne soit respectivement de 6,4,2 et 1 ; ceci détermine les constantes a et b. La quatrième colonne de l'image ne contient que du bruit. La forme spatiale des régions actives est rendue circulaire de diamètre 8 par la convolution de chaque carré avec une gaussienne dont le maximum coïncide avec le centre du carré. Cette convolution est faite avant que le bruit gaussien ne soit ajouté et a donc pour effet de faire varier localement le CNR sans changer le bruit.

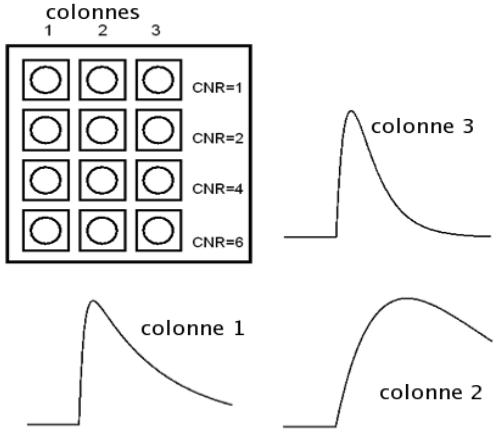


FIG. 3.2 – Construction de l'image hyper-spectrale

Les résultats obtenus sont présentés à la Figure 3.5. Tous les cercles sont différenciés sauf les cercles de la première troisième colonne pour la première ligne (les deux spectres qui se ressemblent le plus pour le rapport contraste sur bruit le plus faible). Il reste aussi deux pixels isolés d'affectation erronée. Enfin les cercles sont quelque peu déformés. Les résultats obtenus sont radicalement différents de ceux obtenus par Witcher *et al.* [76]. En effet, mis à part quelques pixels affectés à des zones isolées après l'étape de détection de frontières, nos zones sont régulières dans le sens où elles n'ont que peu ou pas de trous correspondants à des erreurs. Ceci résulte directement de la structure de notre procédure combinant une méthode de lissage par noyau (non linéaire) à une méthode de segmentation. Ceci a pour effet d'une part de produire une segmentation dans laquelle les distances dans l'image ont leur rôle à jouer et d'autre part de réduire assez significativement le bruit pour pouvoir différencier certaines zones regroupées à tort dans la démarche de Witcher *et al.* Notre algorithme tire pleinement parti de la taille des zones homogènes et de la régularité des frontières.

Nous tenons aussi à remarquer que l'image utilisée n'est pas vraiment constante par régions. En effet, chaque carré étant convolé avec une gaussienne d'intensité maximale au centre du carré, les cercles détectés ne sont pas séparés du reste de l'image par une frontière matérialisée par une rupture, mais correspondent à une zone d'inflexion de la gaussienne utilisée lors de la convolution. Ceci nous permet de conclure à une certaine robustesse de l'algorithme utilisé.

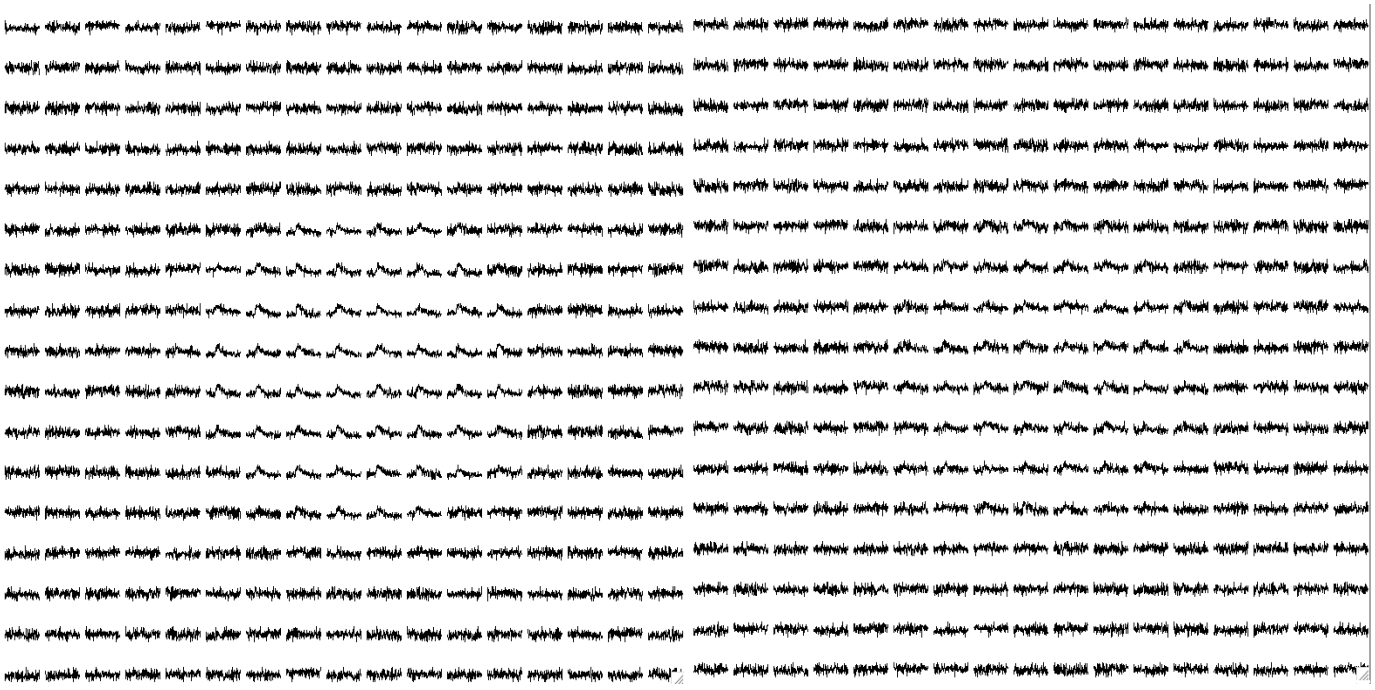


FIG. 3.3 – image hyper-spectrale construite colonne 3 ligne 3 et 2

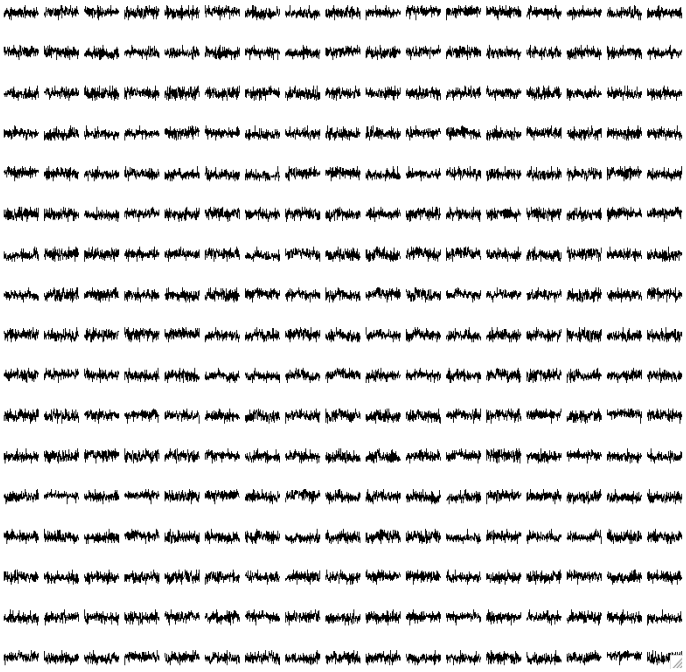


FIG. 3.4 – image hyper-spectrale construite colonne 3 ligne 1 (CNR=1)

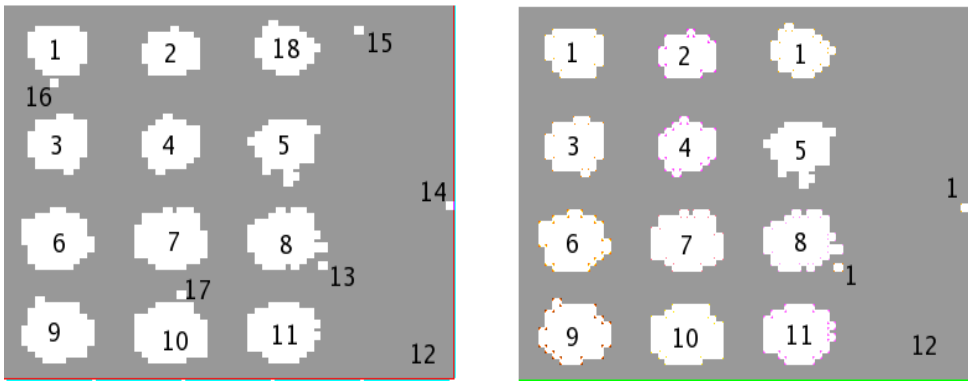


FIG. 3.5 – Carte des zones identifiées avant (à gauche : 18 zones) et après (à droite 12 zones) le regroupement de zones par minimisation d'un critère pénalisé.

3.6 Conclusion et perspectives

Nous avons présenté un algorithme de segmentation d'images hyper-spectrales. Cet algorithme tire parti de l'homogénéité de certaines zones et de la régularité des frontières les délimitant pour séparer de manière convaincante les différentes régions. Notons que la phase d'estimation donne lieu à des problèmes théoriques difficiles liés à la dépendance. La partie algorithmique de recherche de frontière de poids maximum mériterait d'être améliorée en utilisant par exemple des algorithmes du type minimum ratio weight cycle (voir Jermyn [43]) et des algorithmes plus rapides de formation de zones à partir des frontières (voir par exemple [39]). Enfin, l'hypothèse sur la structure de covariance des données est trop restrictive et on peut imaginer la possibilité d'estimer, en même temps que les moyennes, les covariances des spectres par lissage adaptatif.

3.7 Quelques calculs

Choix des paramètres de AWS

Nous allons décrire dans les sous-sections suivantes deux manières de déterminer les paramètres $(\lambda_i^k)_{i,k}$ introduits dans la Sous-section 3.2.3.

Contrôle des faux positifs

Soit i un pixel fixé, nous cherchons à décrire les quantités liées aux tests faits à l'étape k dans V_i^k. Nous rappelons qu'un rejet de H_{0j} à tort correspond à une fausse détection de frontière (fausse alarme) c'est-à-dire un faux positif pour le pixel $j \in V_i^k$.

On remarque que les espérances des nombres de vrais positifs et de faux positifs sont respectivement égales au nombre moyen de pixels $j \in V_i^k$ sélectionnés à raison :

$$\sum_{j \in V_i^k, H_{0j} \text{ Vrai}} P(H_{0j} \text{ accepté}) = \mathbb{E}\left[|\widehat{V}_i^{k+} \cap A_{m(i)}|\right],$$

et à tort :

$$\sum_{j \in V_i^k, H_{0j} \text{ Faux}} P(H_{0j} \text{ accepté}) = \mathbb{E}\left[|\widehat{V}_i^{k+} \cap A_{m(i)}^c|\right].$$

($A^c_{m(i)}$ est le complémentaire de $A_{m(i)}$ dans $\{1\dots n\}^2$). L'espérance du nombre de positifs est égale au nombre moyen de pixels $j \in V_i^k$ sélectionnés :

$$\sum_{j\in V_i^k} P(H_0 \text{ accepté}) = \mathbb{E}\left[|\widehat{V}_i^{k+}|\right].$$

Polzehl et Sopkoiny dans [59] cherchent à contrôler la probabilité que le pixel j soit déclaré positif à tort. Aussi, leur région de rejet étant bilatérale, ils choisissent $\lambda \approx 3 \approx z_{1-0,05/2}$ ($\Phi^{-1}(\alpha_{ij}) = z_\alpha$). Dans notre cas, en supposant aussi que $T(\Delta^k_{ij})$ est de loi normale centrée réduite sous H_{0j}, on peut choisir $\lambda_i^k = 1,69 = z_{1-0,05}$. Le choix $\lambda_i^k = 2$ donne de bons résultats. Cependant, il est encore plus intéressant, après un certain nombre d'étapes, de faire en sorte que le seuil puisse grandir afin d'espérer que la probabilité que le pixel j soit un faux positif s'approche de zéro lorsque la taille des voisinages estimés grandit. Nous avons donc choisi :

$$\lambda_i^k = \max\{2, \sqrt{2\log(|\widehat{V}_i^{k-1+}|)}\}.$$

Cependant, l'erreur d'estimation de θ_i à l'étape k est liée à une problématique de test multiple ; c'est ce qui nous a amené à chercher à contrôler la proportion de faux positifs.

Contrôle de la proportion de faux positifs

Soit i un pixel fixé, à l'étape k, $|V_i^k|$ tests sont effectués pour l'estimation de θ_i. On note FP_i^k le nombre de faux positifs, VP_i^k le nombre de vrais positifs, FN_i^k le nombre de faux négatifs, et FP_i^k le nombre de faux positifs correspondants. Ces quantités sont résumées dans le [tableau 3.1]

	accepté	rejeté	total
H_0 vrai	VP_i^k	VN_i^k	$\lvert V_i^{k+} \cap A_{m(i)}\rvert$
H_0 faux	FP_i^k	FN_i^k	$\lvert V_i^{k+} \cap A^c_{m(i)}\rvert$
total	$\lvert\widehat{V}_i^{k+}\rvert$	$\lvert V_i^{k+}\rvert - \lvert\widehat{V}_i^{k+}\rvert$	$\lvert V_i^{k+}\rvert$

TAB. 3.1 – Quantités associées aux $|V_i^k|$ tests

Dans ce cadre, l'espérance de la proportion de faux positifs parmi les positifs est notée FDR_i^k (pour False Discovery Rate) et l'espérance de la proportion de vrais négatifs parmi les négatifs est notée FNR_i^k (pour False Non-discovery Rate) :

$$FDR_i^k = \mathbb{E}\left[\frac{FP_i^k}{FP_i^k + VP_i^k} 1_{]0;\infty[}(FP_i^k + VP_i^k)\right], \quad FNR_i^k = \mathbb{E}\left[\frac{FN_i^k}{FN_i^k + VN_i^k} 1_{]0;\infty[}(FN_i^k + VN_i^k)\right].$$

Si l'image est composée de deux zones A_1 et A_2, A_1 contenant i et sur laquelle θ_i ($i \in A_1$) vaut 0 et A_2 sur laquelle θ_j ($j \in A_2$) vaut a, et si les observations ne dépendaient pas des poids w_{ij}^k (ce qui n'est qu'une approximation), on aurait :

$$\mathbb{E}[\hat{\theta}_i^k - \theta_i] = a \times FNR_i^k \quad \text{et donc} \quad \mathbb{E}[\Delta_{ij}^k] = a(FNR_i^k - FNR_j^k) + (\theta_i - \theta_j).$$

Le biais d'estimation est lié à l'erreur FNR_i^k. Etant donné qu'il n'est pas possible de contrôler FNR_i^k mais qu'un bon contrôle de FDR_i^k peut amener un bon contrôle de FNR_i^k (voir[74] ou encore [2]), nous avons cherché à contrôler l'erreur FDR_i^k par une borne $q = 0.01$. Le contrôle du FDR_i^k se fait en utilisant une variante de la procédure de Benjamini, et Hochberg ([9]). En effet, nos données étant stochastiquement dépendantes, l'utilisation des résultats de Benjamini et Yekutieli ([10]) est nécessaire. D'autre part, nous avons utilisé l'interprétation Bayesienne de Storey ([71] [70] [72]) et introduit une information à priori sur la proportion π_i^k de vraies H_{0j}. A chaque étape k, cette information est réactualisée. Le seuil λ_i^k est finalement choisi de la façon suivante :

- Ordonner (dans l'ordre croissant) les p-valeurs $p_j = 1-\Phi(T_{ij}^k),\ \ j \in V_i^k$ ($\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée réduite et $T_{ij}^k = T(\Delta_{ij}^k)$ est défini dans l'algorithme.

- Choisir
$$j_{FDR} = \max\left\{ j :\ p_{(j)} \leq \frac{jq}{|V_i^k|\pi_i^k \sum_{p\in V_i^k} 1/p} \right\}, \quad \lambda_i^k = T_{ij_{FDR}}^k.$$

- A l'étape suivante estimer la proportion a priori d'hypothèses H_{0j} vraies par :
$$\pi_i^k = |\widehat{V}_i^{k-1+}|/|V_i^k|.$$

Si, sous H_0, $\mathbb{E}[\Delta_{ij}^k] = 0$ alors ce type de démarche permet en effet de contrôler le FDR_i^k par q. Il est aussi raisonnable de chercher à diminuer q avec le nombre d'étapes pour espérer obtenir un biais asymptotiquement nul. Les résultats pratiques obtenus ne sont pas bien meilleurs que ceux obtenus avec la méthode décrite au paragraphe précédent. Son coût algorithmique étant élevé, nous avons choisi, dans les applications, de n'utiliser que la première démarche.

Calcul et estimation de $\mathbb{E}_0[M(\Gamma)]$

Sans perte de généralité, admettons que $u_g \in A_l$ et $u_d \in A_r$. Posons $\tilde{A}_m(u) = A_m \cap Vois(u) \setminus \{u_g, u_d\}$, l'ensemble des votants de la zone m privée de $\{u_g, u_d\}$ et $\mu_{mi} = a_m - a_i$ la différence entre la valeur de f dans la zone m et dans la zone i. Nous avons

$$\mathbb{E}[M(u, Vois(u))] = 2P(\hat{w}_{ij} = 0) \sum_{m=1}^{M} \sum_{s\in\tilde{A}_m(u)} P(D_m(s)),$$

avec

$$D_m(s) = \{\ \text{« } s \in A_m \text{ vote pour u »}\ \} = \{\hat{w}_{ik} = \hat{w}_{jk}\},$$

et puisque les votes sont finalement obtenus par le test d'hypothèses fonctionnelles décrit dans la partie 3.2, on peut écrire :

$$D_m = \{T(\xi_m - \xi_g - \mu_{ml}) < \lambda \text{ et } T(\xi_m - \xi_d - \mu_{mr}) > \lambda\} \cup \{\{T(\xi_m - \xi_g - \mu_{ml}) > \lambda \text{ et } T(\xi_m - \xi_d - \mu_{mr}) < \lambda\}\},$$

où les ξ_i pour $i \in \{1 \dots p\}^2$ ont la même loi que ξ, vecteur aléatoire gaussien de dimension N, centré, de matrice de covariance identitée. Les notations

$$\beta_\lambda(\mu) = P(T(\mu + \xi_m) > \lambda) \quad \text{et} \quad \phi_\lambda(\mu_1, \mu_2) = \mathbb{E}[\beta_\lambda(\xi_d - \mu_1)(1 - \beta_\lambda(\xi_g - \mu_2))].$$

nous permettent d'obtenir, par conditionnement par rapport à ξ_g et ξ_d, l'expression de l'espérance :

$$\mathbb{E}[M(u, Vois(u))] = 2P(T(\mu_{gd} + \xi_g - \xi_d) > \lambda) + \sum_{m=1}^{M} |\tilde{A}_m(u)| \left(\phi_\lambda(\mu_{mr}, \mu_{ml}) + \phi_\lambda(\mu_{ml}, \mu_{mr})\right). \tag{3.17}$$

Si il n'y a pas de frontière entre u_g et u_d cela signifie que $l = r$ et par conséquent :

$$\mathbb{E}_0[M(u, Vois(u))] = 2P(\mathcal{N}(0,1) > \sqrt{2}\lambda) + 2\sum_{m=1}^{M} |\tilde{A}_m(u)| \phi_\lambda(\mu_{ml}, \mu_{ml}).$$

Cette espérance est estimée par une méthode de type plug-in :

$$\widehat{E_0}[M(u, Vois(u))] = \sum_{i \in Vois(u) \setminus \{u_g, u_d\}} \left(\phi_{\lambda_i^k}(\hat{\theta}_i^k - \hat{\theta}_{u_g}^k, \hat{\theta}_i^k - \hat{\theta}_{u_g}^k) + \phi_{\lambda_i^k}(\hat{\theta}_i^k - \hat{\theta}_{u_d}^k, \hat{\theta}_i^k - \hat{\theta}_{u_d}^k)\right)$$

$$+2P(\mathcal{N}(0,1) > \sqrt{2}\lambda).$$

(les $\hat{\theta}_i^k$ étant les estimateurs obtenus par l'algorithme AWS.)

Annexe

Annexe A

Généralités sur l'approximation et l'estimation par seuillage

Nous allons rappeler rapidement quelques généralités sur la théorie de l'approximation et son lien avec l'estimation par seuillage. Cette théorie est au coeur de notre vision de la réduction de dimension en classification (cf partie II). Nous voulons seulement donner les idées qui la sous-tendent pour éviter au lecteur qui ne connaîtraît pas cette approche d'avoir à faire un travail de recherche bibliographique. Malgré son importance, nous omettrons donc la plupart des démonstrations. Nous conseillons la lecture de l'article de Candes [19]. Pour un approfondissement, et avant une lecture des articles de Donoho et Johnstone, nous conseillons la lecture du livre de Mallat [53] et en particulier du Chapitre 10. Avant de motiver l'estimation par seuillage, nous allons donner une petite introduction à la théorie de l'approximation. Nous terminerons en donnant le pendant de cette théorie en estimation de matrices ou d'opérateurs.

A.1 Approximation

La théorie de l'approximation cherche entre autre des réponses à la question suivante. Etant donné un espace vectoriel normé $\mathcal{X}$, et un entier n, quel est (s'il existe) le sous espace de $\mathcal{X}_n$ de dimension n qui minimise :

$$\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{X}_n\subset\mathcal{X}} \|x-y\|_{\mathcal{X}}. \tag{A.1}$$

Si cette question n'a pas de réponse simple, peut-on donner une borne supérieure à cette quantité (en fonction de n) ? Notons que le livre de référence en la matière est celui de Pinkus [58]. Un cadre simple et central pour étudier ce problème est celui des sous-espaces de $l^2(\mathbb{N})$.

A.1.1 Quelques sous-espaces de l^2

Ellipsoide l^p. Soient $p > 0$, $a \in R^\infty$ décroissant vers 0 et telle que $a_1 = 1$. On définit l'ellipsoide l^p de rayon R associée à a par

$$\mathcal{E}_{a,p}(R) = \left\{ f \in l^2,\ \sum_{k>0} \left| \frac{f_k}{a_k} \right|^p \leq R^p \right\}. \tag{A.2}$$

Pour de telles ellipsoides, le problème d'approximation a été totalement résolu par Kolmogorov (voir [58]).

Boules wl^p (l^p faible). Elles sont définies par

$$wl^p(R) = \left\{ f \in l^2 \ \ \forall \eta > 0 \ \ \sum_{i>0} \left| \frac{1_{|f_i|>\eta}}{\eta} \right|^p \leq R \right\}. \tag{A.3}$$

Les boules l^p faibles sont particulièrement adaptées à la description de la qualité d'approximation.

Proposition A.1. *Soient $f \in l^2$ et $(f_n^*)_n$ la suite des éléments de f rangés par ordre décroissant de $|f_n|$. Alors $f \in wl^p(R)$ si et seulement si $f_n^* \leq Rn^{-1/p}$. D'autre part si $p < 2$ et si f_N est l'élément de l^2 constitué des N plus grands termes de f, alors il existe R tel que $f \in wl^p(R)$ si et seulement si il existe C tel que*

$$\|f - f_N\|_{l^2} \leq CN^{-s}, \ \ s = 1/p - 1/2.$$

Ellipsoide de Besov. Soient $R > 0$, $p > 0$, $q \in]0, \infty[$ et $s' > (1/p - 1/2)_+$. En fixant $s = s' - (1/p - 1/2)_+$ on définit la boule de Besov (et la norme associée $\| \ \|_{b_{s',p,q}}$) par

$$\mathcal{B}_{s',p,q}(R) = \left\{ f \in l^2,\ \sum_{j\geq 0} \left[2^{js} \left(\sum_{k=2^j}^{2^{j+1}-1} |f_k|^p \right)^{1/p} \right]^q \leq R^q \right\} = \left\{ f \in l^2,\ \ \|f\|_{b_{s',p,q}} \leq R \right\}, \tag{A.4}$$

si $q < \infty$ et

$$\mathcal{B}_{s',p,\infty}(R) = \left\{ f \in l^2,\ \sup_{j\geq 0} 2^{js} \left(\sum_{k=2^j}^{2^{j+1}-1} |f_k|^p \right)^{1/p} \leq R \right\} = \left\{ f \in l^2,\ \ \|f\|_{b_{s',p,\infty}} \leq R \right\}. \tag{A.5}$$

Si $p \leq q$ alors on a $\mathcal{B}_{s',p,p} \subset \mathcal{B}_{s',p,q}$. Pour la théorie de l'approximation des ellipsoides de Besov, voir l'article de De Vore [23].

A.1.2 Espaces de Fonctions liés à ces parties de l_2

Le cadre des sous-espaces de l^2 est lié à celui d'espaces de fonctions grâce à des décompositions dans une base. Ainsi, la base de Fourier permet de lier les espaces de Sobolev à certaines ellipsoides (dites de sobolev) et la base d'ondelette permet de relier les espaces de Besov aux ellipsoides de Besov.

Base d'ondelettes. Nous rappelons une définition de la base d'ondelette. Une base d'ondelette est construite à partir de deux fonctions ψ et ϕ (l'ondelette mère et l'ondelette père) par dilatation-translation :

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^j k}{2^j}\right), \quad \phi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\phi\left(\frac{t-2^j k}{2^j}\right). \tag{A.6}$$

Nous rappelons qu'une ondelette ψ a p moments nuls si

$$\int_{\mathbb{R}} t^k \psi(t)dt = 0, \quad \forall 0 \leq k < p,$$

et qu'elle est de régularité r si elle est à support compact et a r dérivées continues. Il est possible (voir par exemple [21]) de trouver une ondelette de régularité r telle que

$$\left\{(\phi_{J,k})_{k\in\{1,\dots,2^J\}}, (\psi_{j,k})_{j>J,k\in\{1,\dots,2^j\}}\right\},$$

soit une base orthonormée de $L^2[0,1]$. Pour toute fonction f de $L^2[0,1]$ on a alors la décomposition suivante :

$$f = \sum_{k=1}^{2^J} \theta_{J,k}\phi_{Jk} + \sum_{j>J} \theta_{j,k}\psi_{jk},$$

où $\theta_{J,k} = \langle f, \phi_{Jk}\rangle$ sont appelés coefficients d'échelle et $\theta_{j,k} = \langle f, \psi_{jk}\rangle$ (pour $j < J$) coefficients d'ondelette.

Espaces de Besov. Les coefficients d'ondelettes mesurent la régularité globale d'une fonction. Soit $\Delta_h^{(r)} f = \sum_{k=0}^r C_r^k f(t+kh)$ la différence d'ordre r. Le $r$ieme module de régularité de $f \in L^p[0,1]$ est

$$w_{r,p}(f,h) = \|\Delta_h^{(r)} f\|_{L^p[0,1-rh]}.$$

La seminorme de Besov d'indices (s',p,q) est définie pour $r > s'$ par

$$|f|_{B_{s',p,q}} = \left(\int_0^1 \left(\frac{w_{r,p}(f,h)}{h^{s'}}\right)^q \frac{dh}{h}\right)^{1/q}$$

si $q < \infty$ et sinon, par

$$|f|_{B_{s',p,\infty}} = \sup_{0<h<1} \frac{w_{r,p}(f,h)}{h^{s'}}.$$

L'espace de Besov $B_{s',p,q}$ est l'ensemble des fonctions $f \in L^p[0,1]$ telles que $|f|_{B_{s',p,q}} < \infty$. On a par exemple $H^s[0,1] = B[0,1]_{s,\infty,\infty}$.

La proposition qui suit exprime le lien entre les espaces de Besov et les ellipsoides de Besov via la base d'ondelette (voir par exemple [28]).

Proposition A.2. *Etant donnée une base d'ondelette de régularité $r > s'$, $1 \leq p,q \leq \infty$ $f \in L^p$ et $\theta(f)$ le vecteur des coefficients d'ondelette de f, il existe deux constantes $0 < c_1, c_2$ indépendantes de f telles que*

$$c_1\|\theta\|_{b_{s',p,q}} \leq \|f\|_{L^p} + |f|_{B_{s',p,q}} \leq c_2\|\theta\|_{b_{s',p,q}}. \tag{A.7}$$

A.2 Estimation par seuillage et approximation

Cette section paraphrase une petite partie de l'introduction de Candes [19], où l'on trouve toutes les références aux articles de Donoho et Johnstone.

Estimateur par seuillage. Supposons que $\mathcal{F}(R)$ est une boule de $\mathbb{R}^n$ ou de l^2 telle que celles que l'on a défini dans la Section 1. On observe un vecteur $\theta \in \mathcal{F}(R)$ dans un bruit gaussien centré :

$$Y_i = \theta_i + \sigma\epsilon_i \;\; i = 1, \ldots, n \;\; \sigma > 0 \;\; \epsilon \rightsquigarrow \mathcal{N}(0, I_n). \tag{A.8}$$

La constante d'échelle σ est supposée connue et l'on veut estimer θ par $\hat{\theta}$ tout en garantissant une petite erreur quadratique $MSE_n(\theta, \hat{\theta}) = E[\|\theta - \hat{\theta}\|_2^2]$. Si l'on cherche à construire un estimateur (diagonal) de la forme $\hat{\theta}^w$ où $\hat{\theta}_i^w = w_i Y_i$, le meilleurs des estimateurs pour le risque MSE_n est celui qui fait le meilleurs compromis entre biais d'estimation et variance d'estimation, c'est-à-dire l'estimateur $\hat{\theta}^*$ obtenu à partir des poids w_k^* minimisant :

$$(1 + w_k)^2\theta_k^2 + w_k^2\sigma^2.$$

Si maintenant on se restreint aux estimateurs pour lesquels $w \in \{0, 1\}^n$, on montre que l'estimateur optimal est $\hat{\theta}^I$ donné par

$$\hat{\theta}_i^I = w_i Y_i \quad \text{et } w_k = 1 \text{ si } |\theta_i| \geq \sigma \text{ et } 0 \text{ sinon .}$$

On montre facilement que cet estimateur a un risque de l'ordre de celui de $\hat{\theta}^*$:

$$MSE(\theta, \hat{\theta}^*) \leq MSE(\theta, \hat{\theta}^I) \leq 2MSE(\theta, \hat{\theta}^*),$$

et que donc la restriction correspondant à choisir $w \in \{0, 1\}^n$ ne coûte quasiment rien. Par ailleurs un simple calcul donne

$$MSE(\theta, \hat{\theta}^I) = \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2).$$

L'estimateur $\hat{\theta}^I$ est construit à partir de la connaissance de θ. La question à laquelle il faut alors répondre est : peut-on construire, à partir des observations seulement, un estimateur par seuillage (c'est-à-dire avec $w \in \{0, 1\}^n$), qui ait un risque proche de $MSE(\theta, \hat{\theta}^I)$? La résponse est donnée par les travaux de Donoho et Johnstone. En notant $\hat{\theta}^\lambda$ l'estimateur défini par

$$\hat{\theta}_i^\lambda = Y_i 1_{|Y_i| \geq \lambda},$$

Donoho et Johnstone montrent que si $\lambda = \sigma\sqrt{\log(n)}$, alors :

$$MSE(\theta, \hat{\theta}^\lambda) \leq (2\log(n) + 1)\left(\sigma^2 + MSE(\theta, \hat{\theta}^I)\right). \tag{A.9}$$

Notons que cet estimateur peut aussi s'interpréter comme un estimateur obtenu par maximum de vraisemblance pénalisé (cf [5]).

Lien avec l'approximation. Le risque $MSE(\theta, \hat{\theta}^I)$ determine donc avec σ l'erreur finale. Il est donné par

$$MSE(\theta, \hat{\theta}^I) = \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2) = N(\sigma)\sigma^2 + \sum_{k>N(\sigma)} |\theta|_{(k)}^2,$$

où $N(\sigma)$ est le nombre de coefficients de θ supérieurs ou égaux en valeur absolue à σ et $|\theta|_{(k)}$ est la valeur absolue du $k$ième plus grand coefficient de θ en valeur absolue. Le terme $N(\sigma)\sigma^2$ est un terme de variance et $e_{N(\sigma)}(\theta) = \sum_{k>N(\sigma)} |\theta|_{(k)}$ est un terme de biais mesurant la qualité d'approximation de θ par ses plus grandes coordonnées :

$$e_B(\theta) = \|\theta - \theta_B\|^2,$$

(où θ_B est l'approximation de θ obtenue en ne gardant que les B plus grands coefficients de θ). Ainsi, la Proposition A.1 nous indique que si $\theta \in wl_p$, alors $e_k(\theta) = O(k^{-1/p+1/2})$. Dans ce cas un équilibrage du biais et de la variance dans $MSE(\theta, \hat{\theta}^I)$ permettent d'obtenir en notant $1/m = 1/p - 1/2$,

$$MSE(\theta, \hat{\theta}^I) = O(\sigma^{\frac{4m}{2m+1}}).$$

En conclusion, l'erreur d'estimation par seuillage de θ au vu de Y (A.8) est directement liée à l'erreur d'approximation $e_B(\theta)$. Dans un cadre minimax, c'est la pire des erreurs d'estimation qui est considérée. Elle est directement liée à l'erreur d'approximation donnée par (A.1).

A.3 Cas des opérateurs

Dans le cas où $\mathcal{X}$ est un espace d'opérateurs, la théorie de l'approximation est moins aboutie. D'une manière générale, si $\mathcal{X}$ composé d'opérateurs pour lesquels on sait trouver une base qui les diagonalise tous, on peut ramener le problème au cas où $\mathcal{X}$ est un espace de séquences. Dans le cas contraire, on peut diviser l'erreur d'approximation en deux parties : l'une concernant la distance des opérateurs à une sous classe d'opérateurs diagonaux dans une base connue, et l'autre concernant l'erreur d'approximation des valeurs propres. Pour la deuxième partie de l'erreur, on peut utiliser la théorie de l'approximation sur les séquences. Il reste alors à étudier la première partie de l'erreur d'approximation.

Quasi diagonalisation d'opérateurs avec les ondeletttes. Notons pour cette partie l'existence de résultats sur la base d'ondelette. Meyer [55] donne un résultat (qui est présenté comme une étape à la démonstration du théorème $T(1)$) permettant de mesurer la décroissance des coefficients hors diagonale d'un opérateur de "Calderon-Zygmund" (Classe d'opérateurs) dans la base d'ondelette. Mallat et Al. [54] donnent un résultat du même type mais dans une base de cosinus locaux et avec un formalisme différent (définition de la classe d'opérateur envisagé dans le domaine de fourier à l'aide des opérateurs intégraux de Fourier). L'originalité des travaux de ces derniers réside dans l'interprétation statistique de la classe d'opérateurs. En effet, Mallat et Al. [54] montrent que ces classes d'opérateurs presque diagonaux dans la base d'ondelette contiennent les opérateurs de covariances associés à des processus localement stationnaires.

Approximation et estimation d'opérateurs. Dans le cas des opérateurs, le même type de raisonnement que celui fait dans la Section « Estimation par seuillage et approximation » peut être fait. Une estimation par seuillage peut être vue à deux niveaux : le seuillage des coefficients hors diagonal et le seuillage des coefficients dans la diagonale. Ces deux types de seuillages correspondent aux deux approximations successives que nous avons décrites dans la section précédente. Nous ne décrivons pas plus le problème correspondant et renvoyons le lecteur au rapport technique de Donoho et ses collaborateurs [26] pour une introduction détaillée à ces problèmes.

Annexe B

Mesure gaussienne sur les espaces de Banach

Nous allons introduire quelques rappels et notations sur les mesures fortement intégrables dans un espace de Banach. La description que nous donnons est un peu rapide pour comprendre le fond de la théorie des variables aléatoires gaussiennes infinies dimensionnelles sans les avoir étudiées au préalable. Cependant, elle permet de définir les espaces naturels en jeu (pour γ une mesure gaussienne, l'espace $L_2(\gamma)$ et l'espace $H(\gamma)$ dit auto-reproduisant) dans les problèmes qui sont les nôtres. Nous pourrons ainsi définir à travers un formalisme rigoureux les objets limites qui régissent nos problèmes en grande dimension. Nous insistons sur le fait que la compréhension des quantités en jeu n'est pas simplement utile en dimension infinie, mais permet de comprendre quelles sont les quantités qui, dans la pratique pour la classification en grande dimension, jouent un rôle important. Ainsi le lecteur qui n'est pas familier avec les espaces de Banach pourra remplacer dans une lecture grossière les espaces de Banach par $\mathbb{R}^p$. Ce qu'il s'agira de comprendre c'est l'importance de la norme de RKHS (Reproducing Kernel Hilbert Space) pour la mesure de la séparation des données et les conditions d'orthogonalité de mesures gaussiennes.

Dans ce qui va suivre nous supposerons que $\mathcal{X}$ est un espace de Banach séparable et nous noterons $\langle\ ,\ \rangle_{\mathcal{X}^*,\mathcal{X}}$ le produit de dualité entre $\mathcal{X}$ et $\mathcal{X}^*$ son dual topologique[1]. Lorsque $\mathcal{X}$ sera un espace de Hilbert, son produit scalaire sera noté $\langle\ ,\ \rangle_{\mathcal{X}}$. Nous rappelons que dans un espace de Banach $\mathcal{X}$, une application linéaire A de $\mathcal{X}$ dans lui même (ou opérateur) est dite bornée si il existe une constante C telle que pour tout $x \in \mathcal{X}$, $\|Ax\|_{\mathcal{X}} \leq C\|x\|_{\mathcal{X}}$. Si un opérateur A est non borné sur $\mathcal{X}$, $\mathcal{D}(A)$ est son ensemble de définition (i.e l'ensemble des éléments x de $\mathcal{X}$ pour lesquels $\|Ax\|_{\mathcal{X}} < \infty$). Si $\mathcal{X}$ est un espace de Hilbert, un opérateur A est dit auto-adjoint lorsque pour tout $(x, y) \in \mathcal{D}(A)$, $\langle Ax, y\rangle_{\mathcal{X}} = \langle x, Ay\rangle_{\mathcal{X}}$. Notons que puisque nous supposerons que les courbes observées appartiennent à un espace de Banach séparale $\mathcal{X}$, cette mesure peut être étendue de manière unique en une mesure de Radon ($\mathcal{X}$ étant en outre polonais, i.e. espace métrique complet séparable). Nous rappelons ici quelques résultats sur les mesures gaussiennes de Radon et nous renvoyons le lecteur au livre de Bogachev [15] pour un exposé complet sur le domaine. Il s'agit de garder à l'esprit que l'idée directrice de ce formalisme, et notre motivation quant à son utilisation, est d'avoir un cadre limite pour la classification dans $\mathbb{R}^p$ quand p tend vers l'infini.

[1] L'ensemble des formes linéaires continues sur $\mathcal{X}$

Dans toute la suite γ sera soit une mesure de probabilité centrée[2] sur la tribu borélienne $\mathcal{B}(\mathcal{X})$ de $\mathcal{X}$, de carré fortement intégrable, c'est-à-dire telle que $\int_{\mathcal{X}} \|x\|^2 \gamma(dx) < \infty$, soit une mesure de probabilité image d'une telle mesure par une translation sur $\mathcal{X}$. Sous cette hypothèse, lorsque γ est centrée, l'intégrale de Bochner $\int_{\mathcal{X}} f(x)x\gamma(dx)$ est parfaitement définie pour tout $f \in L^2(\mathcal{X}, \gamma)$ et définit un élément de $\mathcal{X}$. Ainsi l'opérateur de covariance R de la mesure γ de moyenne μ sera défini comme étant l'opérateur de $\mathcal{X}^*$ dans $\mathcal{X}$ donné par :

$$Rf = \int_{\mathcal{X}} f(x-\mu)(x-\mu)\gamma(dx). \tag{B.1}$$

Une telle mesure γ est dite gaussienne lorsque pour tout $f \in \mathcal{X}^*$ $\gamma \circ f^{-1}$ est une mesure gaussienne sur $\mathbb{R}$. Si γ est une mesure gaussienne centrée sur $\mathcal{X}$, alors pour $h \in \mathcal{X}$, l'application qui à $x \in \mathcal{X}$ associe $h + x$ (translation sur $\mathcal{X}$) est mesurable et la mesure image de γ par cette application est une mesure gaussienne de moyenne h. Nous rappelons que dans le formalisme associé aux mesures infinies dimensionnelles, deux espaces jouent un rôle fondamental.

1. Le RKHS (Reproducing Kernel Hilbert Space) encore appelé espace de Cameron Martin (sera défini par la suite) et noté $H(\gamma)$.
2. L'espace $L^2(\mathcal{X}, \gamma)$ (nous écrirons aussi $L^2(\gamma)$) des fonctions mesurables par rapport à la mesure γ et de carré intégrable par rapport à cette mesure.

Si γ est de moyenne μ, nous noterons S l'opérateur linéaire continu de $L^2(\mathcal{X}, \gamma)$ dans $\mathcal{X}$ définit par

$$\forall f \in L_2(\mathcal{X}, \gamma) \;\; Sf = \int_{\mathcal{X}} (y-\mu) f(y-\mu)\gamma(dy). \tag{B.2}$$

Nous avons (cf [48]) :

$$\|Sf\|_{\mathcal{X}} \leq \left(\int_{\mathcal{X}} \|y-\mu\|_{\mathcal{X}}^2 \gamma(dy) \right)^{1/2} \|f(.-\mu)\|_{L^2(\mathcal{X},\gamma)}, \tag{B.3}$$

et par conséquent S est continu. En identifiant $L^2(\mathcal{X}, \gamma)$ à son dual, il est possible de définir l'opérateur adjoint S^* associé par :

$$\forall x^* \in \mathcal{X}^*, \; f \in L_2(\mathcal{X}, \gamma), \quad \langle S^* x^*, f \rangle_{L^2(\mathcal{X},\gamma)} = \langle x^*, Sf \rangle_{\mathcal{X},\mathcal{X}^*} = \int_{\mathcal{X}} x^*(y) f(y) \gamma(dy). \tag{B.4}$$

Il est clair que S^* est l'application qui à tout élément de $\mathcal{X}^*$ associe sa classe d'équivalence dans $L^2(\mathcal{X}, \gamma)$. Nous noterons $\mathcal{X}_\gamma^*$ la fermeture dans $L^2(\mathcal{X}, \gamma)$ de $S^*(\mathcal{X}^*)$; nous obtenons un espace de Hilbert séparable. De plus, $Ker(S) = (\mathcal{X}_\gamma^*)^\perp$ et $Im(S) = S(\mathcal{X}_\gamma^*)$. En effet, $Sf = 0$ équivaut à $\langle x^*, Sf \rangle = 0$ pour tout $x^* \in \mathcal{X}^*$, et donc à $f \in S^*(\mathcal{X}^*)^\perp$. Il suffit ensuite de décomposer $L^2(\mathcal{X}, \gamma)$ en $L^2(\mathcal{X}, \gamma) = \mathcal{X}_\gamma^* \oplus (\mathcal{X}_\gamma^*)^\perp$ pour voir que $Im(S) = S(\mathcal{X}_\gamma^*)$. D'autre part, on définit $H(\gamma)$ par $H(\gamma) = S(\mathcal{X}_\gamma^*)$. Il est clair que $H(\gamma)$ est en correspondance bijective linéaire avec $\mathcal{X}_\gamma^*$, de sorte que l'on peut transporter la structure hilbertienne de $\mathcal{X}_\gamma^*$ sur l'espace $H(\gamma)$. Du fait de (B.3), l'inclusion naturelle de $H(\gamma)$ dans $\mathcal{X}$, notée j, est continue. En notant j^* l'opérateur $S \circ S^*$ de $\mathcal{X}^*$ dans $H(\gamma)$, le mécanisme des applications introduites se résume avec le diagramme suivant :

$$\mathcal{X}^* \xrightarrow{j^*} H(\gamma) \xhookrightarrow{j} \mathcal{X}\,.$$

[2] autrement dit pour tout $f \in \mathcal{X}^*$, $\gamma \circ f^{-1}$ est une mesure centrée sur $\mathbb{R}$

Il permettra de résoudre les problèmes statistiques que nous avons envisagés. Notons que dans le cas triviale où $\mathcal{X} = \mathbb{R}^p$, et γ est la mesure gaussienne centrée réduite sur $\mathbb{R}^p$, alors $\mathcal{X}^*_\gamma$ est l'espace engendré par p variables aléatoires réelles gaussiennes indépendantes identiquement distribuées selon une loi normale centrée réduite. Notre utilisation des objets introduits sera principalement basée sur les remarques des paragraphes suivant.

Si γ_μ n'est pas centrée mais de moyenne μ, alors on se ramène à une mesure centrée γ par l'application d'une translation pour définir les espaces $\mathcal{X}^*_{\gamma_\mu}$ et $H(\gamma_\mu)$. Ainsi, S et S^* sont définis par le biais de γ, $\mathcal{X}^*_{\gamma_\mu}$ est la fermeture dans $L^2(\gamma_\mu)$ de l'ensemble des classes d'équivalence des éléments de $\{f - \int_{\mathcal{X}} f(x)\gamma(dx), \quad f \in X^*\}$, et $H(\gamma_\mu) = S(\mathcal{X}^*_{\gamma_\mu})$. Soit maintenant γ une mesure sur $\mathcal{X}$ non nécessairement centrée. Si $(e_i^*)_i$ est une famille de $\mathcal{X}^*$ telle que $(S^*(e_i^*))_i$ soit une base orthonormée de $\mathcal{X}^*_\gamma$, alors $(j^*(e_i^*))_i$ est une base orthonormée de $H(\gamma)$. On peut en fait associer à toute base orthonormée $(e_i)_i$ de $H(\gamma)$ une base orthonormée $(e_i^*)_i$ de $\mathcal{X}^*_\gamma$, constituée d'éléments de $\mathcal{X}^*$ tels que $e_k = jj^*(e_k^*)$. On définit alors, pour $x \in \mathcal{X}$, l'opérateur linéaire continu de $\mathcal{X}$ dans $\mathcal{X}$ $P_p(x) = \sum_{i=1}^p e_p^*(x)e_p$. C'est un projecteur. Sa restriction à $H(\gamma)$ est la projection orthogonale sur le sous-espace vectoriel $Vect(e_1, \ldots, e_p)$ de $H(\gamma)$ engendré par $(e_i)_{i=1,\ldots,p}$.

Dans la suite, la mesure γ sera supposée gaussienne (La condition $\int_{\mathcal{X}} \|x\|^2\gamma(dx) < \infty$ étant alors vérifiée pour γ une mesure centrée). L'espace de Hilbert $H(\gamma)$ est alors l'espace auto-reproduisant associé au noyau de covariance K_γ de γ définit par

$$\forall x^*, y^* \in \mathcal{X}^*_\gamma \quad K_\gamma(x^*, y^*) = \int_{\mathcal{X}} x^*(z-\mu)y^*(z-\mu)\gamma(dz),$$

où μ est la moyenne de la mesure γ. Nous rappelons que le support de la mesure γ de moyenne μ est $\bar{H}(\gamma) + \mu$ (où $\bar{H}(\gamma)$ est la fermeture dans $\mathcal{X}$ de $H(\gamma)$). Si la mesure centrée associée à γ est fortement intégrable (donc en particulier si γ est une mesure gaussienne), la suite $(P_p)_p$ de projecteur converge fortement γ-presque sûrement vers l'identité de $\mathcal{X}$. Cette propriété des opérateurs de projection illustre une caractéristique essentielle d'une mesure gaussienne (et plus généralement d'une mesure de probabilité fortement intégrable) sur un espace de Banach séparable. Le comportement d'une telle mesure peut se décrire comme limite de ses restrictions à des sous espaces de dimensions finies. Dans toute la suite, $\gamma_{R,m}$ sera une mesure gaussienne de covariance R et de moyenne m. Soient $f \in L^1(\gamma_{R,m})$, $P_p(x) = \sum_{i=1}^p e_p^*(x)e_p$ et $f_p = f \circ P_p$ la restriction de f au sous-espace de $\mathcal{X}$ de dimension p engendré par $(e_1, \ldots, e_p)$. On peut ramener l'intégrale d'une fonction $f \in L^1(\gamma_{R,m})$, à la limite de l'intégrale de f_p par rapport $\gamma_{R_p,m_p} = \gamma \circ P_p^{-1}$ lorsque p tend vers l'infini. En effet

$$\lim_{p\to\infty} \int_{\mathcal{X}} f_p(x)\gamma_{R_p,m_p}(dx) = \lim_{p\to\infty} \int_{\mathcal{X}} f(P_p(x))\gamma_{R,m}(dx),$$

ce qui d'après le théorème de convergence dominée de Lebesgue permet d'écrire :

$$\lim_{p\to\infty} \int_{\mathcal{X}} f_p(x)\gamma_{R_p,m_p}(dx) = \int_{\mathcal{X}} f(x)\gamma_{R,m}(dx).$$

Cette dernière équation exprime en quoi une mesure gaussienne $\gamma_{C,m}$ permettra de décrire le comportement asymptotique du problème de classification dans $\mathbb{R}^p$.

Dans les deux problèmes de classification envisagés dans $(\mathbb{R}^p, \gamma_{C_1^p,\mu_1^p}, \gamma_{C_0^p,\mu_0^p})$ (procédures LDA et QDA) les fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ ($\mathcal{L}_{10}^p(x)$) permettant de décrire les règles de classification sont soit affines soit quadratiques. Nous aurons besoin d'utiliser les objets mathématiques vers lesquels ces fonctions tendent lorsque p tend vers l'infini. C'est le rôle que jouent les espaces $\mathcal{X}^*_{\gamma_{R,m}}$ (pour les application affines) et $E_2(\gamma_{R,m})$ (pour les polynômes de degrés deux). Nous allons donner la définition de $E_2(\gamma_{R,m})$. Nous en donnons la définition dans $\mathcal{X} = \mathbb{R}^p$ et rappelons une méthode permettant de généraliser cette définition dans le cas des espaces de banach séparable.

Définition B.1. *Soit $f \in L_2(\gamma_{R,m})$. L'application f est un polynôme de degrés deux de carré intégrable, s'il existe $(e_i^*)_{i\geq 0}$ une base orthonormée de $\mathcal{X}^*_{\gamma_{R,m}}$, $\alpha = (\alpha_i)_{i\geq 0} \in l^2$, $\beta = (\beta_{i\geq 0})_i \in l^2$, et $c \in \mathbb{R}$ tels que*

$$f = c + \sum_{i\geq 0} \alpha_i e_i^* + \sum_{i\geq 0} \beta_i((e_i^*)^2 - 1).$$

L'ensemble des polynômes de degrés deux de carré intégrable est noté $\mathcal{X}^*_{2,\gamma}$. On définit $E_2(\gamma_{R,m})$ par la décomposition

$$\mathcal{X}^*_{2,\gamma} = \{cte\} \oplus \mathcal{X}^*_{\gamma} \oplus E_2(\gamma_{R,m}), \tag{B.5}$$

où $\{cte\}$ est l'espace des fonctions constantes.

Notons que l'espace $E_2(\gamma_{R,m})$ peut être défini à partir de l'ensemble $HS(H(\gamma_{R,m}))$ des opérateurs Hilbert-Schmidt[3] sur $H(\gamma_{R,m})$. A un opérateur $A \in HS(H(\gamma_{R,m}))$, de valeurs propres $(\lambda_i)_i$ dans une base orthonormale $(e_n)_n$ de $H(\gamma_{R,m})$, est associée la forme quadratique mesurable de moyenne nulle et de carré intégrable

$$q_A^{\gamma_{R,m}} = \sum_{i>0} \lambda_i((e_i'(x))^2 - 1), \tag{B.6}$$

où $(e_n')_n$ est la base de $\mathcal{X}^*_{\gamma_{R,m}}$ associée à $(e_n)_n$. En outre, on a la relation

$$\|q_A^{\gamma_{R,m}}\|^2_{L_2(\gamma_{R,m})} = 2\|A\|^2_{HS(H(\gamma_{R,m}))}. \tag{B.7}$$

Cette relation est la version limite du fait que si $(\xi_i)_{i=1,\dots,p}$ est un vecteur gaussien centré réduit de $\mathbb{R}^p$, alors

$$\mathrm{Var}\left(\sum_{i=1}^p \lambda_i \xi_i^2\right) = 2\|\lambda\|^2_{\mathbb{R}^p}.$$

En dimension finie, si R est de rang plein, on peut aussi écrire

$$q_A^{\gamma_{R,m}}(x) = \langle AR^{-1/2}(x-m), R^{-1/2}(x-m)\rangle_{\mathbb{R}^p} - \sum_{i=1}^n \lambda_i, \tag{B.8}$$

et

$$\|q_A^{\gamma_{R,m}}\|^2_{L_2(\gamma_{R,m})} = 2\|A\|^2_{HS(H(\gamma_{R,m}))} = 2\|R^{-1/2}AR^{1/2}\|^2_{HS(\mathcal{X})}. \tag{B.9}$$

[3]Un opérateur K sur $\mathcal{X}$ un espace de Hilbert séparable est Hilbert Schmidt si pour une base orthonormée $(e_i)_i$ de $\mathcal{X}$, $\sum_i \|Ke_i\|^2_{\mathcal{X}} < \infty$.

Annexe C

Un brin de théorie de l'information : l'inégalité de Kraft

L'objectif de cette annexe est de présenter l'inégalité de Kraft dans le cadre très particulier de la description de la richesse d'un modèle. Elle est motivée par le Chapitre 2 de la troisième partie de ce mémoire.

Soit $\mathcal{M}$ un modèle fini, c'est-à-dire un ensemble fini d'objets quelconques. Chaque objet x de $\mathcal{M}$ est étiqueté (on dit aussi codé) par une suite de 0 et de 1 de longueur finie. Nous appellerons fonction de codage et noterons ϕ l'application qui a un objet $x \in \mathcal{M}$ associe le code correspondant et $l(\phi(x))$ la longueur de ce code (nous noterons $n = \max_x l(x)$ la plus grande de ces longueur). L'application ϕ est supposée injective (deux objets qui ont deux étiquettes différentes sont différents). Nous appellerons code et nous noterons $\phi(\mathcal{M})$ l'image de $\mathcal{M}$ par ϕ. La fonction ϕ détermine donc l'étiquetage des objets du modèle $\mathcal{M}$ et $\phi(\mathcal{M})$ est l'ensemble des étiquettes associées aux objets. Si x est un objet de $\mathcal{M}$, alors son étiquette $\phi(x)$ est une suite finie de 0 et de 1 de longueur $l(x)$, nous la noterons $\phi(x) = (\phi_1(x), \ldots, \phi_{l(x)}(x))$ et nous appellerons préfixes de l'étiquette de x les suites $(\phi_1(x), \ldots, \phi_k(x))$ pour $1 \leq k \leq l(x)$.

On dit qu'un code $\phi(\mathcal{M})$ a la propriété de préfixe si pour tout objet $x \in \mathcal{M}$, il n'existe pas d'autre objet $x' \in \mathcal{M}$ tel que x' soit un préfixe de x. Un tel code est aussi dit instantané et présente l'intérêt pratique suivant. Imaginons que x' soit un objet de longueur $l(x') = 1$ (ou $l(x)$ petit) et que l'on reçoit la concaténation de $\phi(x')$, $\phi(x_1), \ldots, \phi(x_k)$ ($x_1, \ldots, x_k$ k autres éléments de $\mathcal{M}$). Si le code est préfixé, la lecture du premier terme de cette concaténation permet effectivement de décider si celui-ci est x' ou pas. Si le code n'est pas préfixé, il faut lire les n premiers termes de la concaténation. En théorie du codage il est intéressant de trouver des codes qui sont préfixés et qui n'ont pas des étiquettes trop longues (ils sont plus rapides à lire). Le théorème de Kraft (1949) et Mc Millan (1956) donne une condition sur la longueur des étiquettes.

Theoreme C.1 (Kraft, Mc Millan)**.** *Si ϕ est une fonction de codage (injective) alors les longueurs des codes satisfont l'inégalité de Kraft :*

$$\sum_{x \in \mathcal{M}} 2^{-l(\phi(x))} \leq 1. \tag{C.1}$$

De plus s'il existe une fonction de codage ϕ telle l'inégalité de kraft soit vérifiée, alors il existe une fonction de codage $\tilde{\phi}$ préfixée telle que pour tout objet x de $\mathcal{M}$ $l(\phi(x)) = l(\tilde{\phi}(x))$.

La première partie de ce théorème est dûe à Kraft dans le cas de fonction de codage préfixée et à Mc Millan dans le cas général. La deuxième partie est dûe à Kraft.

Application Soit $\mathcal{M}_N$ l'ensemble de modèles défini dans le Chapitre 2 de la partie III. Nous allons construire un code pour coder les éléments de ce modèle. Pour un élément $u \in \mathcal{M}$, il s'agit de coder deux choses :

- la partition récursive diadique associée à u, et
- la valeur des proportions du mélange sur cette partition.

Proposition C.1. *Il existe un code avec une fonction de codage injective ϕ qui vérifie la propriété suivante. Si $u \in \mathcal{M}_N$ est associé à une partition $P(u)$ de taille m alors $l(\phi(u)) = \frac{pen(u)}{\log(2)}$, où*

$$pen(u) = m\left(\frac{3}{2}(K-1)\log N + \frac{4}{3}\log 2\right). \tag{C.2}$$

Démonstration.

- **Codage de la partition.** Tout arbre ayant m feuilles peut être codé en utilisant au plus m bits pour les feuilles, et au plus $\frac{m-1}{3}$ bits pour les noeuds qui ne sont pas des feuilles. Aussi, $\frac{4}{3}m$ bits suffisent pour coder l'arbre entier, ce qui implique la deuxième partie de l'expression de $pen(u)$.
- **codage de la valeur des poids.** Pour une classe donnée $i \in \{1, \ldots, K\}$, et une région donnée R parmi les m régions de la partition, le poids associé à la classe i est discrétisé sur une grille régulière de pas $\frac{1}{N^{1/3}}$. Pour coder ce poids, il faut donc $\log_2\left(N^{\frac{3}{2}}\right)$ bits. Rappelons que la connaissance sur R des poids associés aux $(K-1)$ premières classes implique la connaissance des poids associés à toutes les classes. La première partie de l'expression de $pen(u)$ en découle.

□

De cette proposition et du théorème précédent, on déduit l'inégalité de Kraft (2.7) donnée dans le Chapitre 2 de la partie III.

Bibliographie

[1] F Abramovich, A Antoniadis, T Sapatinas, and B Vidakovic. Optimal testing in a fixed-effects functional analysis of variance model. *International Journal of Wavelets, Multiresolution and Information Processing*, 2004.

[2] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of statistics*, 34, 2006.

[3] T. Anderson and R. Bahadur. Classification into two multivariate normal distributions with different covriance matrices. *Annals of Mathematilcal Statistics*, 33(2) :420–431, 1962.

[4] A Antoniadis, F Abramovich, T Sapatinas, and B Vidakovic. Wavelet methods for testing in functional analysis of variance models. *International Journal on Wavelets and its applications*, 93 :1007–1021, 2004.

[5] A Antoniadis and J Fan. Regularization of wavelet approximations. *JASA*, 96(455) :939, 2001.

[6] J.Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. *Annals of Statistics*, 2006.

[7] Y. Baraud. Non asymptotic minimax rate of testing in signal detection. *Bernoulli*, 8 :577–606, 2002.

[8] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Annals of stats*, 31(1) :225–251, 2003.

[9] Y. Benjamini and Y. Hocheberg. Controlling the false discovery rate :a practical and poweful approach to multiple testing. *Journal of Royal Statistical Society B*, 57 :289–300, 1995.

[10] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4) :1165–1188, 2001.

[11] H. Bensmail and G. Celeux. Regularized gaussian disciminant analysis through aigenvalue decomposition. *JASA*, pages 1743–1748, 1996.

[12] P. Bickel and E. Levina. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6) :989–1010, 2004.

[13] P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 2007.

[14] L. Birgé. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Annales de l'I.H.P. Probabilités et statistiques*, 42(3) :273–325, 2006.

[15] V. I. Bogachev. *Gaussian Measures*. AMS, 1998.

[16] L. Breiman, J. Friedman, Olshen R., and Stone C.J. *Classification and regression trees*. Belmond, CA : Wadsworth, 1983.

[17] F. Bunea, M. Wegkamp, and A. Auguste. Consistent variable selection in high dimensional regression via multiple testing. *Journal of statistical planning and inference*, 136 :4349–4354, 2006.

[18] M. V. Burnashef and Begmatov. On a problem of signal detection leading to stable distribution. *Theory of probability and its applications*, 35(3) :556–560, 1990.

[19] E. Candes. Modern statistical estimation via oracle inequalities. *Acta Numerica*, pages 1–69, 2006.

[20] D Canet, J.C. Boudel, and E Canet Soulas. *La RMN, concepts, méthodes et applications*. Dunod, 2002.

[21] I. Daubechies. *Ten Lectures on Wavelets*, volume 61 of *Lecture Notes*. SIAM, 1992.

[22] B Delyon and A Iouditski. Wavelet estimators global error measures revisited. Technical report, IRIS, Decembre 1993.

[23] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[24] D Donoho. Cart and best-ortho-basis : A connection. *Annals of Statistics*, 25 :1870–1911, 1997.

[25] D Donoho. Wedgelets : Nearly-minimax estimation of edges. *Annals of statistics*, pages 859–897, 1999.

[26] D. Donoho, S. Mallat, and R. Von Sachs. Estimating covariances of locally stationary processes : consistancy of best basis methods. Technical report, Standford, 1996.

[27] D. L. Donoho and I. Johnstone. Minimax risk over lp-balls for lq-error. *Probability Theory and Related Fields*, (99) :277–303, 1994.

[28] D. L. Donoho and I. Johnstone. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3) :879–921, 1998.

[29] L.C. Evans and R.F. Gariepy. *Measure theory and fine property of functions*. CRC press, 1992.

[30] J Fan. Test of significance based on wavelet thresholding and neyman's truncation. *JASA*, 91 :674–688, 1996.

[31] J. Fan and S-K Lin. Test of significance when data are curves. *JASA*, 93 :1007–1021, 1998.

[32] J. Fan and Fan Y. High dimensional classification using features annealed independence rules. Technical report, Princeton University, 2007.

[33] K. Fan. Minimax theorems. *Proc. Nat. Acad. Sc. USA*, 19 :42–47, 1953.

[34] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 :179–188, 1936.

[35] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, 2001.

[36] R. Girard. Segmentation d'images hyperspectrales. *Traitement du signal*, 23 :277–287, 2006.

[37] Ulf Grenander. Stochastic processes and statistical inference. *Arkiv for Matematik*, 1 :195–277, 1950.

[38] A. Gualtierotti. On the stability of signal detection. *IEEE Trans. on Information Theory*, 29(3) :426–433, 1983.

[39] L. Guigues. *Modèles Multi-Echelles pour la Segmentation d'Images*. PhD thesis, IGN, 2003.

[40] G. Hagberg. From magnetic resonance spectroscopy to classification of tumors. a reviewof pattern recognition methods. *NMR in Biomedicine*, 156 :11 :148, 1998.

[41] T. Hastie, A. Buja, and R. Tibshirani. Penalised discriminant analysis. *Annals of Statistics*, 23 :73–102, 1995.

[42] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 :65–70, 1979.

[43] Ian H. Jermyn and H. Ishikawa. Globally optimal regions and boundaries as minimum ratio weight cycles. *IEEE Transaction on pattern analysis and machine intelligence*, 23(10), oct 2001.

[44] S. A. Kassam and H. V. Poor. Robust techniques for signal processing : A survey. *Proceedings of the IEEE*, 73(3), 1985.

[45] M Kohler. Nonparametric estimation of piecewise smooth regression functions. Technical report, Stuttgart, 2003.

[46] E. Kolaczyk, J. Junchang, and S Gopal. Multiscale, multigranular statistical image segmentation. *JASA*, 100(472) :1358, December 2005.

[47] Korostelev and Tsybacov. *Minimax Theory of Image Reconstruction*, volume 82 of *Lecture Notes In Statistics*. Springer-Verlag, 1993.

[48] J. Kuelbs. Gaussian measure on banach space. *J. of Func. Anal.*, 5, 1970.

[49] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The annals of Statistics*, 28(5) :1302–1338, 2000.

[50] M. Ledoux. *The concentration of Measure Phenomenon*. AMS, 2001.

[51] V Lepski, O. On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.*, 35(3) :454, 1992.

[52] Q. J. Li. *Estimation of mixture Models*. PhD thesis, Yale university, 1999.

[53] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[54] S Mallat, G Papanicolaou, and Z Zhang. Adaptive covariance estimation of locally stationary processes. *The annals of Statistics*, 26(1) :1–47, 1998.

[55] Y. Meyer. *Ondelettes et opérateurs, tome 2 : Opérateurs de Calderon-Zygmund*. Hermann, 1997.

[56] V.V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, 1975.

[57] H. Philippe, N. Stransky, J.P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array cgh data : from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18) :3413–3422, 2004.

[58] A. Pinkus. *n-Widths in Approximation Theory*. Springer-Verlag, 1985.

[59] J. Polzehl and V. Spokoiny. Adaptive weights smoothing with applications to image restoration. *J.R Stat Soc B*, 62 :335–354, 2000.

[60] J. Polzehl and V. Spokoiny. Vector adaptive weights smoothing with application to mri. *J.R Stat Soc B*, 63 :335–354, 2001.

[61] J. Polzehl and V. Spokoiny. Propagation-separation approach for local likelihood estimation. *Probability Theory and Related Fields*, 135(3) :335–362, 2006.

[62] C.R. Rao and Varadarajan. Discrimination of gaussian processes. *Sankhya : The Indian Journal of Statistics*, 25 :303–330, 1963.

[63] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69 :730–742, 2006.

[64] W. Rudin. *Analyse réelle et complexe*. Dunod, 1987.

[65] W. Rudin. *Analyse fonctionnelle*. Ediscience, 1995.

[66] Shorack. *Probability for Statistitian*. Springer, 2000.

[67] S. Simons. *Minimax and monotonicity*, volume 1693 of *Lecture Notes in Mathematics*. Springer, 1999.

[68] V. Spokoiny. Adaptative hypothesis testing using wavelets. *Annals of Statistics*, 24(6) :2477–2498, december 1996.

[69] E. Stein. *Harmonic Analysis : Real-Variable Methods, Orthogonality, and Oscilatory Integrals*. Princeton University Press, 1993.

[70] J.D. Storey. The positive false discovery rate : A bayesian interpretation and the q-value. *Annals of Statistics*, 31 :2013–2035, 2003.

[71] J.D. Storey, Taylor JE, and Siegmund. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64 :479–498, 2002.

[72] J.D. Storey, J.E. Taylor, and Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates : A unified approach. *Journal of the Royal Statistical Society, Series B*, 66 :187–205, 2004.

[73] Fabien Szablo de edelenyi. *Développement d'une nouvelle approche d'analyse des images spectroscopiques RMN : les images nosologiques*. PhD thesis, UJF, novembre 2001.

[74] J. Taylor, R. Tibshirani, and B. Efron. The miss rate for the analysis of expression data. *Biostatistics*, 6(1) :111–117, 2005.

[75] A. Tsybakov. *Introduction a l'estimation non-parametrique*. Springer, 2004.

[76] B. Whitcher, A.J. Schwarz, H. Barjat, S. Smart, R. Grundy, and M. F. James. Wavelet-based cluster analysis : Data-driven grouping of voxel time-courses with application to perfusion-weighted and pharmacological mri of the rat brain. *NeuroImage*, 24(2) :281–295, 2005.

[77] Y Yang. Minimax nonparametric classification. part i :rates of convergence part ii model selection for adaptation, part i :i model selection for adaptation :rates of convergence part ii model selection for adaptation. *IEEE Trans. on Information Theory*, 1999.

[78] C Zhu, S and A Yuille. Region competition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9), 1996.